QBS 181 Group Project

Mallory Maher

11/6/2021

Read in Libraries

To reproduce these results, the libraries you will need are readxl in order to read in the data, tidyverse for data manipultation, ggplot, and joining the datasets, and reshape2 in order to melt the data for the plots.

Load in the necessary libraries:

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

Read in and Clean Data

First I read in my data. If you are using the csv's on github, change the read_excel to read_csv and change the file name accordingly. Here, I am reading in the data and changing Entity (country or region) and Code (country or region code) to factors. I then combined both datasets into one.

```r
# country data
mentalHealthDF_country <-
  read_excel("Mental health Depression disorder Data.xlsx")
#str(mentalHealthDF_country)

# regional data
mentalHealthDF_region <-
  read_excel("region data.xlsx")
#str(mentalHealthDF_region)

# converting Entity and Code to factor
mentalHealthDF_country$Entity <- as.factor(mentalHealthDF_country$Entity)
mentalHealthDF_country$Code <- as.factor(mentalHealthDF_country$Code)

mentalHealthDF_region$Entity <- as.factor(mentalHealthDF_region$Entity)
mentalHealthDF_region$Code <- as.factor(mentalHealthDF_region$Code)

# combining regional and country level data
mentalHealthDF <- rbind(mentalHealthDF_country, mentalHealthDF_region)
```

Plot Function

Here, I made my own plot function to prevent excessive copy pasting. This plot takes the parameters of the country or region and the data frame that you are using and will return a plot that will give you the average prevalence of the seven mental health disorders included from the years 1990 to 2017.

```r
country_region_plots <- function(entity, entityDF){
  # load in libraries
  library(tidyverse)
  library(ggplot2)
  library(reshape2)

  # selecting Entity/country of choice
  regionDF_entity <- entityDF %>%
    filter(Entity == entity) %>%
    group_by(Year) %>%
    summarise(Alcohol_Use = mean(`Alcohol use disorders (%)`),
              Anxiety = mean(`Anxiety disorders (%)`),
              Bipolar = mean(`Bipolar disorder (%)`),
              Depression = mean(`Depression (%)`),
              Drug_Use = mean(`Drug use disorders (%)`),
              Eating_Disorder = mean(`Eating disorders (%)`),
              Schizophrenia = mean(`Schizophrenia (%)`))

  # melting data for plotting
  entity_df <- melt(regionDF_entity, id.vars = 'Year',
                        variable.name = 'Mental Health Disorder')

  # plotting all mental health disorder data in a line graph from 1990 to 2017
  ggplot(data = entity_df, aes(x = Year, y = value, color = `Mental Health Disorder`)) +
    geom_line(stat = "identity", size = 1) +
    geom_point(aes(y = value)) +
```

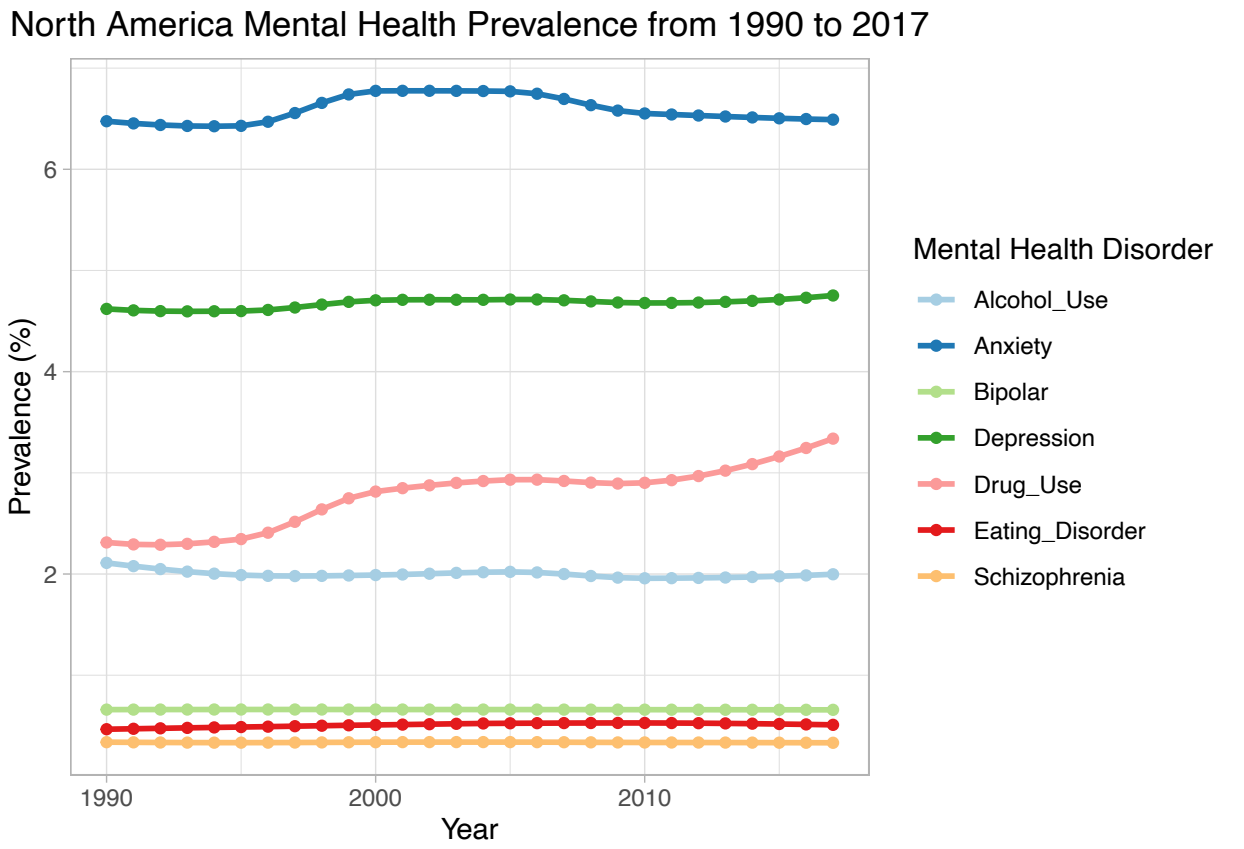
```
    theme_light() +
    labs(title = paste(entity, "Mental Health Prevalence from 1990 to 2017"),
         x = "Year", y = "Prevalence (%)") +
    theme(plot.title = element_text(hjust = 0.5)) +
    scale_color_brewer(palette = "Paired")
}
```

Calling the Average Mental Health Disorder Prevalence Function on Areas of Interest

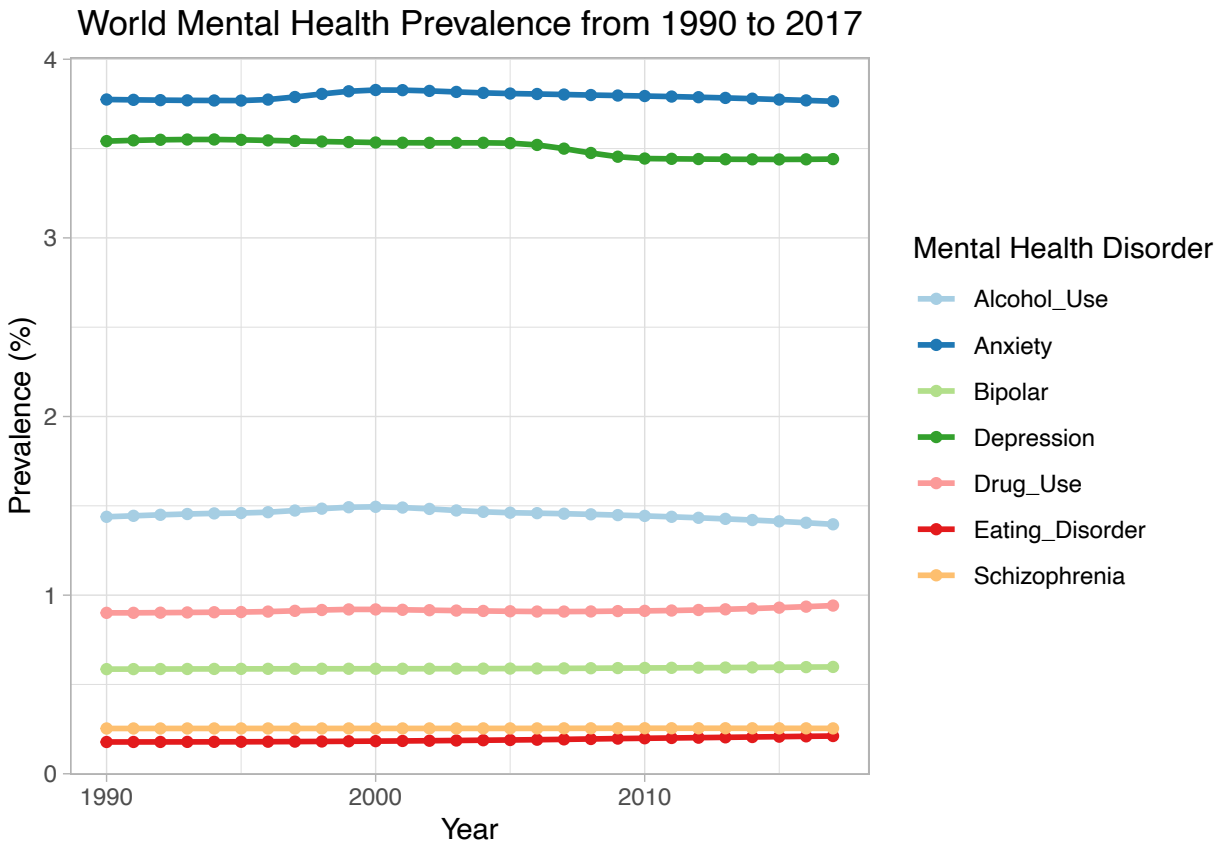
Here, I am calling the plot to give you an example on how to use the function. Don't forget that entity is a string!

Using this function, we can see that Anxiety has the highest prevalence in North America. Depression has the second highest prevalence, which is the disorder that we focused on. You can see that depression is slightly increasing over the years and drug disorder has the highest increase our of all mental health disorders in North America.

```
country_region_plots("North America", mentalHealthDF_region)
```




```
country_region_plots("World", mentalHealthDF_country)
```



Using this function, we can see that Anxiety has the highest prevalence Worldwide, followed closely by depression.

Joining Data - Country

Next, I got the cleaned country data from the other group members (except Emma's because her data was used for a sub-analysis, not this overall analysis). I read in each of the cleaned sheets and then joined them using an inner_join by Entity, Code, and Year. If you were to reproduce this, I would recommend skipping this step and using the full country csv in the github folder.

```
library(readxl)

# loading in everyone's sheets to join them together
libby_sheet <-
  read_excel("Libby_sheet.xlsx")
sean_sheet <-
  read_excel("sean_sheet.xlsx")
carson_sheet <-
  read_excel("carson_sheet.xlsx")

# changed the edited column name for joining
colnames(carson_sheet)[which(names(carson_sheet) == "Country")] <- "Entity"

# joining all data to use multiple variables for upcoming plots and models
full_mentalDF_country <- mentalHealthDF %>%
  inner_join(libby_sheet) %>%
```

```
  inner_join(sean_sheet) %>%
  inner_join(carson_sheet)
```

```
## Joining, by = c("Entity", "Code", "Year")
## Joining, by = c("Entity", "Code", "Year")
## Joining, by = c("Entity", "Code", "Year")
```

```
#write_csv(full_mentalDF_country, "full_mentalHeathDF_country.csv")
```

Joining Data - Region

Next, I got the cleaned regional data from the other group members (except Emma's because her data was used for a sub-analysis, not this overall analysis). I read in each of the cleaned sheets and then joined them using an inner_join by Entity, Code, and Year. If you were to reproduce this, I would recommend skipping this step and using the full region csv in the github folder.

```
library(readxl)
library(tidyverse)

# loading in regional data for joining
sean_sheet2 <-
  read_excel("sean_sheet2.xlsx")
carson_sheet2 <-
  read_excel("carson_sheet2.xlsx")

# joining all region data
full_mentalDF_region <- mentalHealthDF_region %>%
  inner_join(libby_sheet) %>%
  inner_join(sean_sheet2) %>%
  inner_join(carson_sheet2) %>%
  select(-Code)
```

```
## Joining, by = c("Entity", "Code", "Year")
## Joining, by = c("Entity", "Code", "Year")
## Joining, by = c("Entity", "Code", "Year")
```

```
#write_csv(full_mentalDF_region, "full_mentalHeathDF_region.csv")
```

Depression Prevalence by Gender and Age Plot

Here, I made my own plot function to prevent excessive copy pasting. This plot takes the parameters of the country or region and the data frame that you are using and will return a plot that will give you the prevalence of depression by each age range from the years 1990 to 2017 as a line graph and the prevalence of males versus females as a density graph.

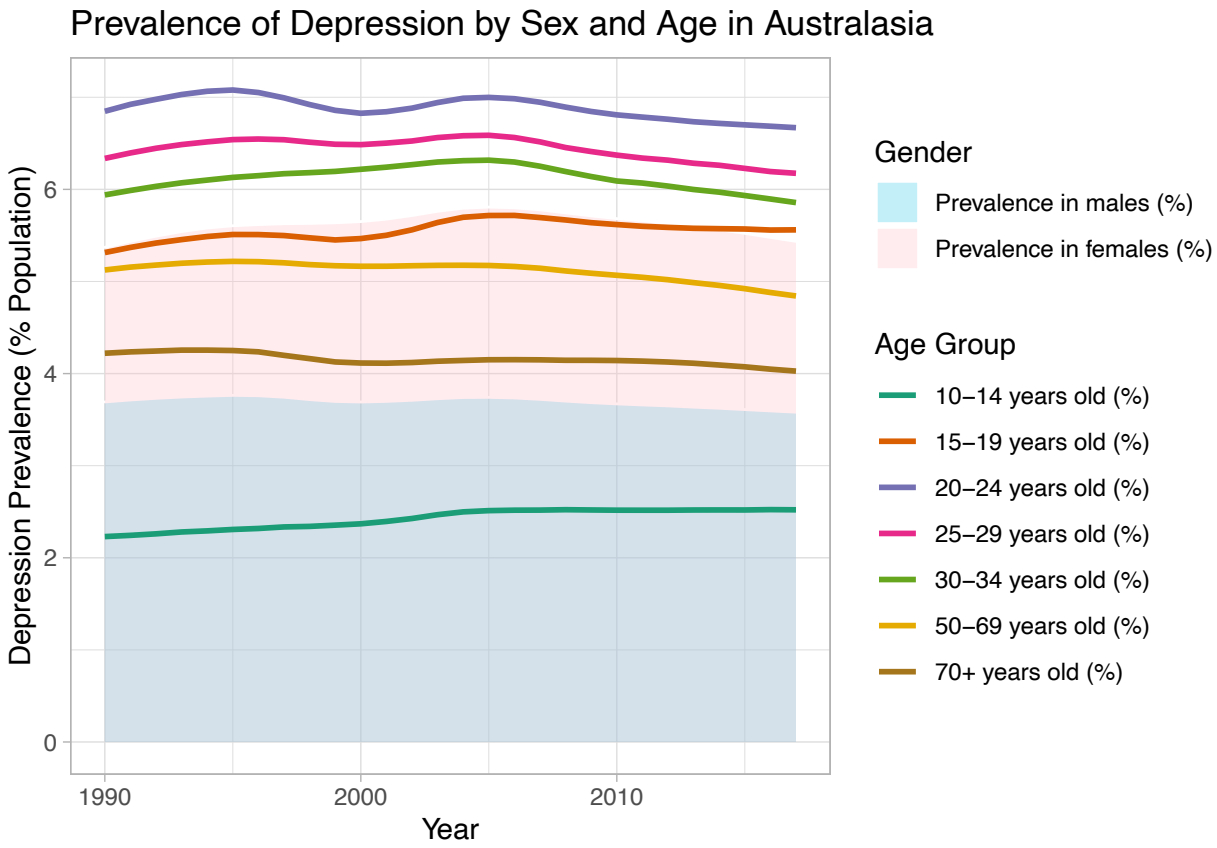
```
age_gender_plots <- function(entity, dataframe){
  # selecting desired entity/country age data
  plotDF_age <- dataframe %>%
    filter(Entity == entity) %>%
```

```
    select(Year, `10-14 years old (%)`,
           `15-19 years old (%)`,
           `20-24 years old (%)`,
           `25-29 years old (%)`,
           `30-34 years old (%)`,
           `50-69 years old (%)`,
           `70+ years old (%)`)

  # selecting desired entity/country sex data
  plotDF_gender <- dataframe %>%
    filter(Entity == entity) %>%
    select(Year, `Prevalence in males (%)`,
           `Prevalence in females (%)`)

  # melting data for plotting
  plotDF2_age <- melt(plotDF_age, id.vars = 'Year')
  plotDF2_gender <- melt(plotDF_gender, id.vars = 'Year')

  # plotting density plot of gender with line graph of age
  ggplot() +
    geom_density(aes(x = Year, y = value, group = variable,
                     fill = variable), color = "white",
                 data = plotDF2_gender, stat = "identity", alpha=0.3) +
    geom_line(aes(x = Year, y = value, group = variable, color = variable),
              data = plotDF2_age, size = 1) +
    #geom_point(aes(x = Year, y = value, color = variable), data = plotDF2_age) +
    labs(y = "Depression Prevalence (% Population)",
         title = paste("Prevalence of Depression by Sex and Age in", entity),
         fill = "Gender", color = "Age Group") +
    theme_light() +
    scale_fill_manual(values = c("#34c9eb", "pink")) +
    scale_color_brewer(palette = "Dark2")
}
```

Calling the Depression Prevalence by Age and Gender Function on Areas of Interest

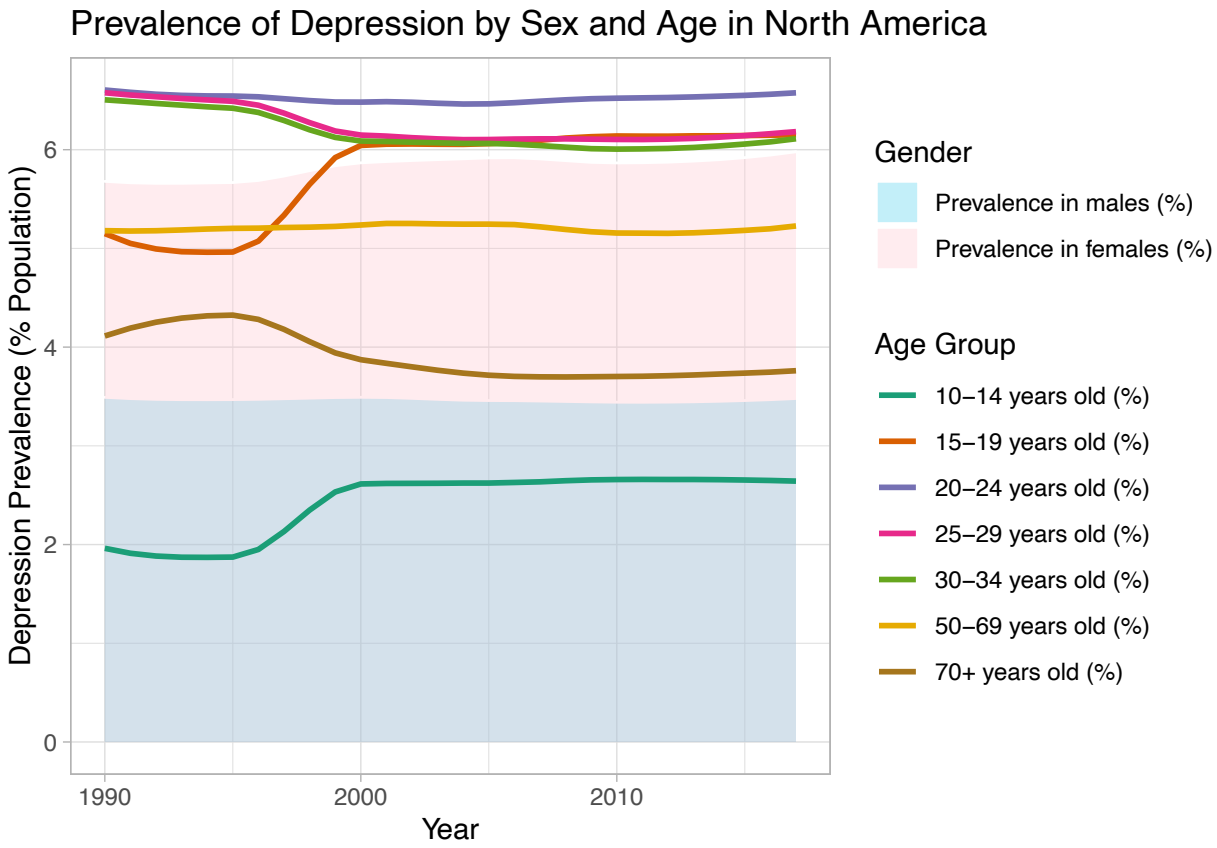
Here, I am calling the plot to give you an example on how to use the function and comparing Australasia, North America, and Greenland. Don't forget that entity is a string!

```
age_gender_plots("Australasia", full_mentalDF_region)
```



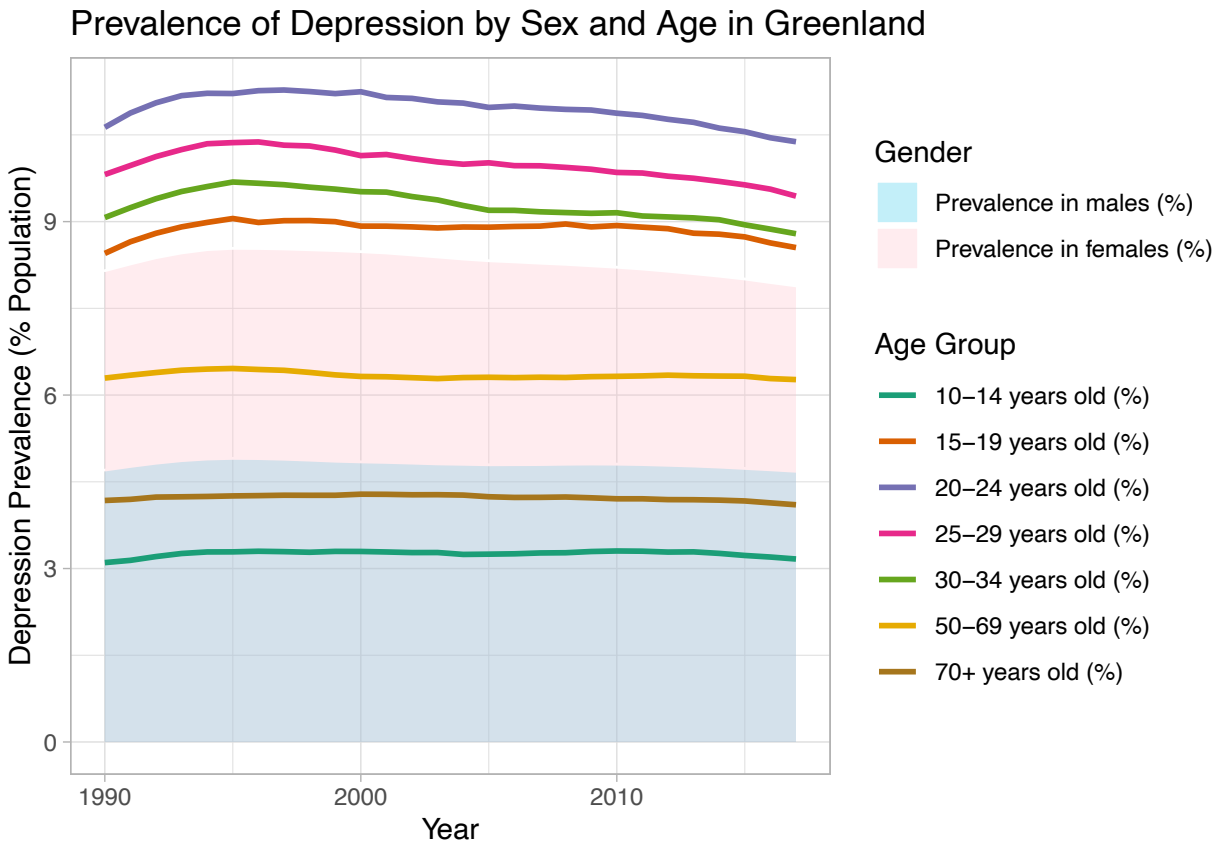
Here, we can see that the younger age groups in Australaisa have higher rates of depression prevalence (more on this in the age documentation). We can also see that the prevalence in females is 1.5% higher than males.

```
age_gender_plots("North America", full_mentalDF_region)
```



Here, we can see that the younger age groups in North America have higher rates of depression prevalence (more on this in the age documentation). We can also see that the prevalence in females is about 2.0% higher than males.

```
age_gender_plots("Greenland", full_mentalDF_country)
```



We noticed on our map plots that Greenland seemed to be an outlier so I decided to look at Greenland specifically. Here, we can see that the y-axis ranges from 0 to 10 whereas in the Australasia and North America plots the axis ranged from 0 to 6.5 or 7. This tells us that Greenland has very high depression prevalence. We hypothesized that this could be from the cold weather and low levels of sunlight. Both of these could lead to seasonal depression. We also noticed that the younger age groups have higher depression prevalence in Greenland, following the trends from North America. We wanted to do this subanalysis on Greenland because this outlier may be pulling up the depression prevalence in North America. The prevalence in females is about 3.5% higher than males. This is a higher disparity than both North America and Australasia.

Linear Model - All countries

We wanted to see which mental health disorders are associated with depression prevalence across all countries, so I made a linear model to assess the coefficients and how other disorders increase or decrease average depression prevalence. What I found was that Depression prevalence increases as Anxiety, Alcohol Use Disorder, Drug Use Disorders, and Eating Disorder prevalence increases and Depression prevalence decreases when Bipolar Disorder and Schizophrenia prevalence increase. All disorders were statistically significant in this model predicting depression.

```
# building linear model - All Countries
mod1 <- lm(`Depression (%)` ~ `Alcohol use disorders (%)` +
   `Anxiety disorders (%)` + `Bipolar disorder (%)` +`Drug use disorders (%)` +
   `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_country)

# calling linear model
summary(mod1)
```

```
##
## Call:
## lm(formula = `Depression (%)` ~ `Alcohol use disorders (%)` +
##     `Anxiety disorders (%)` + `Bipolar disorder (%)` + `Drug use disorders (%)` +
##     `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_country)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16303 -0.39327 -0.06636  0.34808  2.34781
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.471215   0.083534   41.55  < 2e-16 ***
## `Alcohol use disorders (%)`  0.118160   0.009868   11.97  < 2e-16 ***
## `Anxiety disorders (%)`      0.256465   0.012048   21.29  < 2e-16 ***
## `Bipolar disorder (%)`      -1.359996   0.087003  -15.63  < 2e-16 ***
## `Drug use disorders (%)`     0.307762   0.023462   13.12  < 2e-16 ***
## `Eating disorders (%)`       0.786532   0.114162    6.89 6.22e-12 ***
## `Schizophrenia (%)`         -3.164608   0.312311  -10.13  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6026 on 5509 degrees of freedom
## Multiple R-squared:  0.192,  Adjusted R-squared:  0.1911
## F-statistic: 218.2 on 6 and 5509 DF,  p-value: < 2.2e-16
```

Linear Regression Model - United States

We then wanted to see which mental health disorders are associated with depression prevalence in just the United States, so I made another linear model to assess the coefficients and how other disorders increase or decrease average depression prevalence. What I found was that the only disorder that was statistically significant in predicting Depression was Drug Use.

```
# building linear model - United States
mod2 <- lm(`Depression (%)` ~ `Alcohol use disorders (%)` +
   `Anxiety disorders (%)` + `Bipolar disorder (%)` +`Drug use disorders (%)` +
   `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_country[full_mentalDF_country$Enti

# calling linear model
summary(mod2)
```

```
##
## Call:
## lm(formula = `Depression (%)` ~ `Alcohol use disorders (%)` +
##     `Anxiety disorders (%)` + `Bipolar disorder (%)` + `Drug use disorders (%)` +
##     `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_country[full_mentalDF_country$
##     "United States", ])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0057222 -0.0021495 -0.0002826  0.0021886  0.0065375
##
## Coefficients:
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.43900    5.46903   0.446 0.660186
## `Alcohol use disorders (%)`   0.14554    0.08857   1.643 0.115232
## `Anxiety disorders (%)`       0.05332    0.06048   0.882 0.388008
## `Bipolar disorder (%)`        0.80053    7.90110   0.101 0.920259
## `Drug use disorders (%)`      0.16596    0.03921   4.233 0.000372 ***
## `Eating disorders (%)`        0.12084    0.26903   0.449 0.657901
## `Schizophrenia (%)`           1.74741    2.83071   0.617 0.543670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003858 on 21 degrees of freedom
## Multiple R-squared:  0.9965, Adjusted R-squared:  0.9955
## F-statistic:  1000 on 6 and 21 DF,  p-value: < 2.2e-16
```

Linear Regression Model - North America

I wanted to take it one level up from the United States and focus on North America like we did in our presentation, so I made another linear model. What I found was that the only disorders that were statistically significant in predicting Depression were Drug Use and Bipolar Disorder. Depression prevalence increased Drug Use Disorder prevalence increased and decreased when Eating Disorder prevalence increased.

```
# building linear model - North America
mod3 <- lm(`Depression (%)` ~ `Alcohol use disorders (%)` +
   `Anxiety disorders (%)` + `Bipolar disorder (%)` +`Drug use disorders (%)` +
   `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_region[full_mentalDF_region$Entity

# calling linear model
summary(mod3)
```

```
##
## Call:
## lm(formula = `Depression (%)` ~ `Alcohol use disorders (%)` +
##     `Anxiety disorders (%)` + `Bipolar disorder (%)` + `Drug use disorders (%)` +
##     `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_region[full_mentalDF_region$E
##     "North America", ])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0047222 -0.0014617 -0.0002645  0.0017945  0.0052084
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.37058    4.22698  -0.324   0.7490
## `Alcohol use disorders (%)`   0.04719    0.06430   0.734   0.4711
## `Anxiety disorders (%)`       0.05568    0.06410   0.869   0.3949
## `Bipolar disorder (%)`        6.82984    5.96709   1.145   0.2653
## `Drug use disorders (%)`      0.18507    0.02835   6.529 1.82e-06 ***
## `Eating disorders (%)`       -0.58263    0.15574  -3.741   0.0012 **
## `Schizophrenia (%)`           2.55877    2.52068   1.015   0.3216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.003055 on 21 degrees of freedom
## Multiple R-squared:  0.9967, Adjusted R-squared:  0.9958
## F-statistic:  1064 on 6 and 21 DF,  p-value: < 2.2e-16
```

Linear Regression Model - Sub-Saharan Africa

After North America, I wanted to compare my findings to Sub-Saharan Africa like we did in our presentation. What I found was that the disorders that were statistically significant in predicting Depression were Alcohol Use, Bipolar Disorder, Drug Use, and Schizophrenia. Depression prevalence increased as Alcohol Use and Bipolar Disorder prevalence increased and decreased when Drug Use and Schizophrenia prevalence increased.

```
# building linear model - Sub-Saharan Africa
mod4 <- lm(`Depression (%)` ~ `Alcohol use disorders (%)` +
   `Anxiety disorders (%)` + `Bipolar disorder (%)` +`Drug use disorders (%)` +
   `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_region[full_mentalDF_region$Entity

# calling linear model
summary(mod4)
```

```
##
## Call:
## lm(formula = `Depression (%)` ~ `Alcohol use disorders (%)` +
##     `Anxiety disorders (%)` + `Bipolar disorder (%)` + `Drug use disorders (%)` +
##     `Eating disorders (%)` + `Schizophrenia (%)`, data = full_mentalDF_region[full_mentalDF_region$E
##     "Sub-Saharan Africa", ])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0033785 -0.0008551  0.0002989  0.0009319  0.0029952
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.6960     1.9668   0.862 0.398261
## `Alcohol use disorders (%)`    0.4758     0.1362   3.492 0.002172 **
## `Anxiety disorders (%)`       -1.7359     1.0390  -1.671 0.109592
## `Bipolar disorder (%)`        23.8600     4.0284   5.923 7.04e-06 ***
## `Drug use disorders (%)`      -1.5148     0.3559  -4.256 0.000352 ***
## `Eating disorders (%)`         2.1026     2.8455   0.739 0.468120
## `Schizophrenia (%)`          -43.0282     8.4717  -5.079 4.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001799 on 21 degrees of freedom
## Multiple R-squared:  0.9926, Adjusted R-squared:  0.9905
## F-statistic: 471.8 on 6 and 21 DF,  p-value: < 2.2e-16
```

Tableau:
**NOTE: THE DASHBOARD IS ON GITHUB – file: DepressionPrevalenceDashboard.xlsx

First, connect the full mental health dataset by country to a new Tableau worksheet under the heading “Connections”. Once you connect the sheet, change the column name “Entity” to “Country”. This is done because Tableau recognizes the variable “Country” as the actual country that is being listed in the column but would not recognize the country values under a column named “Entity” as actual countries. In order to make choropleth plots, Tableau must understand that those values listed in the column are actual countries so that it can map the appropriate values for each country. To confirm that Tableau is connected to the dataset, when you open a new sheet you should see a list of the variables (Code, Country, Year etc.) on the left hand side under the “Tables” heading. Now, we can start to build the choropleth plots.

First, drag the variable “Country” to the Columns area and the variable “Number of People with Depression” to the Rows area. Your sheet should look like a very basic bar chart after this step. To convert the bar chart to a choropleth plot, click on the “Show Me” icon in the top right of the sheet and select the “maps” option (should be in the middle of the second row of options). Now, the sheet should contain a map of the entire world with the color of each country representing the number of people with depression. We chose the “Sunrise-Sunset Diverging” color palette and set the appropriate start and end counts as seen in the legends on the plots in the dashboard. Next, drag the “Year” variable to the Filters box and then click “Show Filter” so that we can start to build the animation. Finally, click the “Format” tab at the top of your machine and then “Animations” to set up an animation for the plot that can be shown over a time period (1990 – 2017 for our data). Make sure to click “Simultaneous” under the “Style” heading in Animations tab that pops up so that you can drag the Year filter back and forth over the time period and see how the color changes represents the number of people with depression changing over time. To get specific plots by region or country, highlight the region you are interested in and Tableau will automatically gray out the other countries not in the region, allowing you to set a specific range of counts of number of people with depression for each region or country you are interested in.

Project

Libby Czarniak

11/2/2021

Libraries

Throughout the analysis, we're going to need the readxl, tidyverse, dplyr, ggplot2 and ggpubr packages, so we can load all of those in now.

```
##load in all packages
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
library(dplyr)
library(ggplot2)
library(ggpubr)
```

Loading and Cleaning Data

We're going to load in the data sheet with the information on depression of prevalence by age groups and then view it. We'll also make a raw copy of the data. If you're reading in the data from GitHub, simply change the read_excel function to read_csv and change the file name and path as needed.

```
##load in the data
depression_age <- read_excel("~/Documents/data wrangling/Mental health Depression disorder Data 2.xlsx"
    sheet = "prevalence-of-depression-by-age")
View(depression_age)
##make a copy of the raw data
depression_age_raw <- depression_age
```

First, let's look at the head of the data and the structure to get an idea of what we're working with and make sure that the column types are what we need.

```
##preliminary view of the data
head(depression_age)
```

```
## # A tibble: 6 x 13
##   Entity      Code   Year `10-14 years old ~ `15-19 years old~ `20-24 years old~
##   <chr>       <chr> <dbl>              <dbl>             <dbl>              <dbl>
## 1 Afghanistan AFG    1990               1.59              3.46               4.42
## 2 Afghanistan AFG    1991               1.59              3.45               4.43
## 3 Afghanistan AFG    1992               1.58              3.43               4.45
## 4 Afghanistan AFG    1993               1.58              3.42               4.46
## 5 Afghanistan AFG    1994               1.57              3.43               4.46
## 6 Afghanistan AFG    1995               1.57              3.42               4.46
## # ... with 7 more variables: 25-29 years old (%) <dbl>,
## #   30-34 years old (%) <dbl>, 50-69 years old (%) <dbl>,
## #   15-49 years old (%) <dbl>, 70+ years old (%) <dbl>, All ages (%) <dbl>,
## #   Age-standardized (%) <dbl>
```

```
str(depression_age)
```

```
## tibble [6,048 x 13] (S3: tbl_df/tbl/data.frame)
##  $ Entity              : chr [1:6048] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Code                : chr [1:6048] "AFG" "AFG" "AFG" "AFG" ...
##  $ Year                : num [1:6048] 1990 1991 1992 1993 1994 ...
##  $ 10-14 years old (%) : num [1:6048] 1.59 1.59 1.58 1.58 1.57 ...
##  $ 15-19 years old (%) : num [1:6048] 3.46 3.45 3.43 3.42 3.43 ...
##  $ 20-24 years old (%) : num [1:6048] 4.42 4.43 4.45 4.46 4.46 ...
##  $ 25-29 years old (%) : num [1:6048] 5.18 5.18 5.16 5.15 5.15 ...
##  $ 30-34 years old (%) : num [1:6048] 5.8 5.81 5.83 5.85 5.85 ...
##  $ 50-69 years old (%) : num [1:6048] 5.92 5.93 5.95 5.97 5.98 ...
##  $ 15-49 years old (%) : num [1:6048] 4.94 4.9 4.84 4.81 4.84 ...
##  $ 70+ years old (%)   : num [1:6048] 5.2 5.19 5.18 5.17 5.16 ...
##  $ All ages (%)        : num [1:6048] 3.22 3.2 3.16 3.12 3.08 ...
##  $ Age-standardized (%): num [1:6048] 4.07 4.08 4.09 4.1 4.1 ...
```

```
summary(depression_age)
```

```
##     Entity              Code                Year      10-14 years old (%)
##  Length:6048        Length:6048        Min.   :1990   Min.   :0.7103
##  Class :character   Class :character   1st Qu.:1997   1st Qu.:1.0681
##  Mode  :character   Mode  :character   Median :2004   Median :1.2625
##                                        Mean   :2004   Mean   :1.3773
##                                        3rd Qu.:2010   3rd Qu.:1.5871
##                                        Max.   :2017   Max.   :3.3033
##  15-19 years old (%) 20-24 years old (%) 25-29 years old (%)
##  Min.   :1.498       Min.   : 1.719      Min.   : 1.946
##  1st Qu.:2.537       1st Qu.: 3.069      1st Qu.: 3.170
##  Median :2.900       Median : 3.514      Median : 3.718
##  Mean   :3.074       Mean   : 3.781      Mean   : 3.909
##  3rd Qu.:3.408       3rd Qu.: 4.261      3rd Qu.: 4.476
##  Max.   :9.053       Max.   :11.276      Max.   :10.378
##  30-34 years old (%) 50-69 years old (%) 15-49 years old (%) 70+ years old (%)
##  Min.   :2.231       Min.   :3.248       Min.   :2.177       Min.   : 3.249
```

```
##  1st Qu.:3.243       1st Qu.:4.861       1st Qu.:3.310       1st Qu.: 4.899
##  Median :3.971       Median :5.631       Median :3.839       Median : 5.959
##  Mean   :4.076       Mean   :5.672       Mean   :4.014       Mean   : 6.156
##  3rd Qu.:4.733       3rd Qu.:6.409       3rd Qu.:4.587       3rd Qu.: 7.357
##  Max.   :9.685       Max.   :9.778       Max.   :9.626       Max.   :11.532
##   All ages (%)   Age-standardized (%)
##  Min.   :1.806   Min.   :2.140
##  1st Qu.:2.601   1st Qu.:2.976
##  Median :2.996   Median :3.499
##  Mean   :3.265   Mean   :3.499
##  3rd Qu.:3.758   3rd Qu.:3.923
##  Max.   :6.990   Max.   :6.603
```

Based on the preliminary look at the data, we're working with an Entity and Code column that represent the country. These should both be factors so that we can look at meaningful differences between the countries later on. The prevalences (the age group columns) are all a numeric data type because they're percentages, so we can leave those as is. We have 10 columns corresponding to age groups. Lastly, we notice that we don't have any missing values for prevalence, so all of the countries have a measure of prevalence for each age group.

Let's convert the Entity and Code columns both to factors.

```
##change the country-related columns to factors
depression_age$Entity <- as.factor(depression_age$Entity)
depression_age$Code <- as.factor(depression_age$Code)

##look at structure of dataset to make sure they are factors
str(depression_age)
```

```
## tibble [6,048 x 13] (S3: tbl_df/tbl/data.frame)
##  $ Entity              : Factor w/ 216 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Code                : Factor w/ 199 levels "AFG","AGO","ALB",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Year                : num [1:6048] 1990 1991 1992 1993 1994 ...
##  $ 10-14 years old (%) : num [1:6048] 1.59 1.59 1.58 1.58 1.57 ...
##  $ 15-19 years old (%) : num [1:6048] 3.46 3.45 3.43 3.42 3.43 ...
##  $ 20-24 years old (%) : num [1:6048] 4.42 4.43 4.45 4.46 4.46 ...
##  $ 25-29 years old (%) : num [1:6048] 5.18 5.18 5.16 5.15 5.15 ...
##  $ 30-34 years old (%) : num [1:6048] 5.8 5.81 5.83 5.85 5.85 ...
##  $ 50-69 years old (%) : num [1:6048] 5.92 5.93 5.95 5.97 5.98 ...
##  $ 15-49 years old (%) : num [1:6048] 4.94 4.9 4.84 4.81 4.84 ...
##  $ 70+ years old (%)   : num [1:6048] 5.2 5.19 5.18 5.17 5.16 ...
##  $ All ages (%)        : num [1:6048] 3.22 3.2 3.16 3.12 3.08 ...
##  $ Age-standardized (%): num [1:6048] 4.07 4.08 4.09 4.1 4.1 ...
```

Now the Entity and Code columns are all factors, which will allow us to make better use of them later on.

We also notice that the column names for the different age groups are going to be really inconvenient to use in our analysis, so we'll change them to something that's easier to use. For each age group column, we'll convert it to words instead of numbers and remove spaces and the (%). For example, let's take 10-14 years old (%) and change it to ten_to_fourteen. We'll follow the same format for the remainder of the age group columns. Let's use capital letters for the first letters in the all ages and standardized age columns instead of underscores.

```
##change the names of the columns with the age groups
colnames(depression_age)[colnames(depression_age) == 'All ages (%)'] <- 'AllAges'
colnames(depression_age)[colnames(depression_age) == '10-14 years old (%)'] <- 'ten_to_fourteen'
colnames(depression_age)[colnames(depression_age) == '15-19 years old (%)'] <- 'fifteen_to_nineteen'
colnames(depression_age)[colnames(depression_age) == '20-24 years old (%)'] <- 'twenty_to_twentyFour'
colnames(depression_age)[colnames(depression_age) == '25-29 years old (%)'] <- 'twentyFive_to_twentyNin
colnames(depression_age)[colnames(depression_age) == '30-34 years old (%)'] <- 'thirty_to_thirtyFour'
colnames(depression_age)[colnames(depression_age) == '15-49 years old (%)'] <- 'fifteen_to_fortyNine'
colnames(depression_age)[colnames(depression_age) == '50-69 years old (%)'] <- 'fifty_to_sixtyNine'
colnames(depression_age)[colnames(depression_age) == '70+ years old (%)'] <- 'seventy_plus'
colnames(depression_age)[colnames(depression_age) == 'Age-standardized (%)'] <- 'StandardizedAge'

##print out column names to make sure they're changed
colnames(depression_age)
```

```
##  [1] "Entity"                   "Code"
##  [3] "Year"                     "ten_to_fourteen"
##  [5] "fifteen_to_nineteen"      "twenty_to_twentyFour"
##  [7] "twentyFive_to_twentyNine" "thirty_to_thirtyFour"
##  [9] "fifty_to_sixtyNine"       "fifteen_to_fortyNine"
## [11] "seventy_plus"             "AllAges"
## [13] "StandardizedAge"
```

Analysis: Comparing Prevalence between Age Groups across Regions

Now that our data is in the format that we need it in we can start our analysis. We're going to keep all of the columns for now, but we'll revisit some of them later on and decide whether they're going to be useful for the analysis.

Before bringing the data into R, I did some preliminary cleaning and noticed that our data also contains regions of the world rather than just countries in the Entity column. If we look back at the structure of the data, we notice that the Entity column has 216 factors– which is a lot of countries to look at. Therefore, we're going to focus on just regions for now and then pull in country data later on if it'll be necessary or beneficial. This macro-type of analysis will not only make the analysis a little easier but will also give us a more general view of how depression affects different parts of the world.

Let's take a look at what we have in the Entity column and then make a subset of our dataframe with just the regions.

```
##print out all countries in the entity column
unique(depression_age$Entity)
```

```
##   [1] Afghanistan                    Albania
##   [3] Algeria                        American Samoa
##   [5] Andorra                        Angola
##   [7] Antigua and Barbuda            Argentina
##   [9] Armenia                        Australasia
##  [11] Australia                      Austria
##  [13] Azerbaijan                     Bahamas
##  [15] Bahrain                        Bangladesh
##  [17] Barbados                       Belarus
##  [19] Belgium                        Belize
##  [21] Benin                          Bermuda
```

```
##  [23] Bhutan                          Bolivia
##  [25] Bosnia and Herzegovina          Botswana
##  [27] Brazil                          Brunei
##  [29] Bulgaria                        Burkina Faso
##  [31] Burundi                         Cambodia
##  [33] Cameroon                        Canada
##  [35] Cape Verde                      Central African Republic
##  [37] Central Asia                    Central Europe
##  [39] Central Sub-Saharan Africa      Chad
##  [41] Chile                           China
##  [43] Colombia                        Comoros
##  [45] Congo                           Costa Rica
##  [47] Cote d'Ivoire                   Croatia
##  [49] Cuba                            Cyprus
##  [51] Czech Republic                  Democratic Republic of Congo
##  [53] Denmark                         Djibouti
##  [55] Dominica                        Dominican Republic
##  [57] East Asia                       Eastern Europe
##  [59] Eastern Sub-Saharan Africa      Ecuador
##  [61] Egypt                           El Salvador
##  [63] England                         Equatorial Guinea
##  [65] Eritrea                         Estonia
##  [67] Ethiopia                        Fiji
##  [69] Finland                         France
##  [71] Gabon                           Gambia
##  [73] Georgia                         Germany
##  [75] Ghana                           Greece
##  [77] Greenland                       Grenada
##  [79] Guam                            Guatemala
##  [81] Guinea                          Guinea-Bissau
##  [83] Guyana                          Haiti
##  [85] Honduras                        Hungary
##  [87] Iceland                         India
##  [89] Indonesia                       Iran
##  [91] Iraq                            Ireland
##  [93] Israel                          Italy
##  [95] Jamaica                         Japan
##  [97] Jordan                          Kazakhstan
##  [99] Kenya                           Kiribati
## [101] Kuwait                          Kyrgyzstan
## [103] Laos                            Latin America and Caribbean
## [105] Latvia                          Lebanon
## [107] Lesotho                         Liberia
## [109] Libya                           Lithuania
## [111] Luxembourg                      Macedonia
## [113] Madagascar                      Malawi
## [115] Malaysia                        Maldives
## [117] Mali                            Malta
## [119] Marshall Islands                Mauritania
## [121] Mauritius                       Mexico
## [123] Micronesia (country)            Moldova
## [125] Mongolia                        Montenegro
## [127] Morocco                         Mozambique
## [129] Myanmar                         Namibia
```

```
## [131] Nepal                              Netherlands
## [133] New Zealand                        Nicaragua
## [135] Niger                              Nigeria
## [137] North Africa and Middle East       North America
## [139] North Korea                        Northern Mariana Islands
## [141] Norway                             Oceania
## [143] Oman                               Pakistan
## [145] Palestine                          Panama
## [147] Papua New Guinea                   Paraguay
## [149] Peru                               Philippines
## [151] Poland                             Portugal
## [153] Puerto Rico                        Qatar
## [155] Romania                            Russia
## [157] Rwanda                             Saint Lucia
## [159] Saint Vincent and the Grenadines   Samoa
## [161] Sao Tome and Principe              Saudi Arabia
## [163] Scotland                           Senegal
## [165] Serbia                             Seychelles
## [167] Sierra Leone                       Singapore
## [169] Slovakia                           Slovenia
## [171] Solomon Islands                    Somalia
## [173] South Africa                       South Asia
## [175] South Korea                        South Sudan
## [177] Southeast Asia                     Southern Sub-Saharan Africa
## [179] Spain                              Sri Lanka
## [181] Sub-Saharan Africa                 Sudan
## [183] Suriname                           Swaziland
## [185] Sweden                             Switzerland
## [187] Syria                              Taiwan
## [189] Tajikistan                         Tanzania
## [191] Thailand                           Timor
## [193] Togo                               Tonga
## [195] Trinidad and Tobago                Tunisia
## [197] Turkey                             Turkmenistan
## [199] Uganda                             Ukraine
## [201] United Arab Emirates               United Kingdom
## [203] United States                      United States Virgin Islands
## [205] Uruguay                            Uzbekistan
## [207] Vanuatu                            Venezuela
## [209] Vietnam                            Wales
## [211] Western Europe                     Western Sub-Saharan Africa
## [213] World                              Yemen
## [215] Zambia                             Zimbabwe
## 216 Levels: Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```

We have a really long list of countries! Let's pick out just the regions, and here we're going to treat the following as our regions of the world: Australasia, Central Asia, Central Europe, Central Sub-Saharan Africa, East Asia, Eastern Europe, Eastern Sub-Saharan Africa, Latin America and Caribbean, North Africa and Middle East, North America, Oceania, Southeast Asia, Southern Sub-Saharan Africa, Sub-Saharan Africa, Western Sub-Saharan Africa, and Western Europe.

We'll use these regions to make a new dataframe, and we want to filter the original dataframe to include only rows corresponding to these regions.

```r
##use filter to select only rows with region data
depression_age_regions <- filter(depression_age, Entity %in% c("Australasia", "Central Asia", "Central E

##turn our filtered data into a dataframe
depression_age_regions <- as.data.frame(depression_age_regions)
##look at a summary of our dataframe
summary(depression_age_regions)
```

```
##                         Entity         Code          Year      ten_to_fourteen
##  Australasia               : 28   AFG    :  0   Min.   :1990   Min.   :0.8637
##  Central Asia              : 28   AGO    :  0   1st Qu.:1997   1st Qu.:1.0544
##  Central Europe            : 28   ALB    :  0   Median :2004   Median :1.2647
##  Central Sub-Saharan Africa: 28   AND    :  0   Mean   :2004   Mean   :1.4004
##  East Asia                 : 28   ARE    :  0   3rd Qu.:2010   3rd Qu.:1.5896
##  Eastern Europe            : 28   (Other):  0   Max.   :2017   Max.   :2.6600
##  (Other)                   :280   NA's   :448
##  fifteen_to_nineteen twenty_to_twentyFour twentyFive_to_twentyNine
##  Min.   :1.963       Min.   :2.413        Min.   :2.546
##  1st Qu.:2.621       1st Qu.:3.311        1st Qu.:3.310
##  Median :2.966       Median :3.775        Median :3.977
##  Mean   :3.235       Mean   :4.006        Mean   :4.096
##  3rd Qu.:3.581       3rd Qu.:4.442        3rd Qu.:4.765
##  Max.   :6.151       Max.   :7.079        Max.   :6.587
##
##  thirty_to_thirtyFour fifty_to_sixtyNine fifteen_to_fortyNine  seventy_plus
##  Min.   :2.591        Min.   :4.217      Min.   :2.737        Min.   :3.698
##  1st Qu.:3.383        1st Qu.:5.203      1st Qu.:3.448        1st Qu.:5.322
##  Median :4.196        Median :5.886      Median :4.067        Median :6.346
##  Mean   :4.233        Mean   :5.893      Mean   :4.192        Mean   :6.502
##  3rd Qu.:4.938        3rd Qu.:6.703      3rd Qu.:4.814        3rd Qu.:8.114
##  Max.   :6.505        Max.   :7.278      Max.   :6.440        Max.   :9.034
##
##     AllAges      StandardizedAge
##  Min.   :2.301   Min.   :2.439
##  1st Qu.:2.716   1st Qu.:3.171
##  Median :3.005   Median :3.799
##  Mean   :3.403   Mean   :3.652
##  3rd Qu.:4.209   3rd Qu.:3.973
##  Max.   :5.029   Max.   :4.795
##
```

Just from the summary of our dataframe above, we can start to pick out some useful pieces of information about how prevalence of depression differs across age groups. For example, the 10-14 age group has the lowest average prevalence of depression among all of the age groups, and the 70+ age group has the highest average prevalence of depression among all of the age groups. The 70+ group also has some pretty high data points- including 8.114% and 9.034%.

We also notice that there are 3 age groups that are distinct from the rest: 15-49, all ages, and standardized age. I am going to exclude all ages and standardized age from the remainder of the analysis because we're more interested in comparing specific age groups to each other. Although we could compare age groups to the aggregate of all ages, I prefer to focus on how age groups compare directly. Also, the 15-49 age group seems to be making up for the fact that we're missing ages 35-49, but it's also double-counting ages 15-34. Therefore, I am also going to exclude that column from the rest of the analysis and focus on the remainder of the age groups.

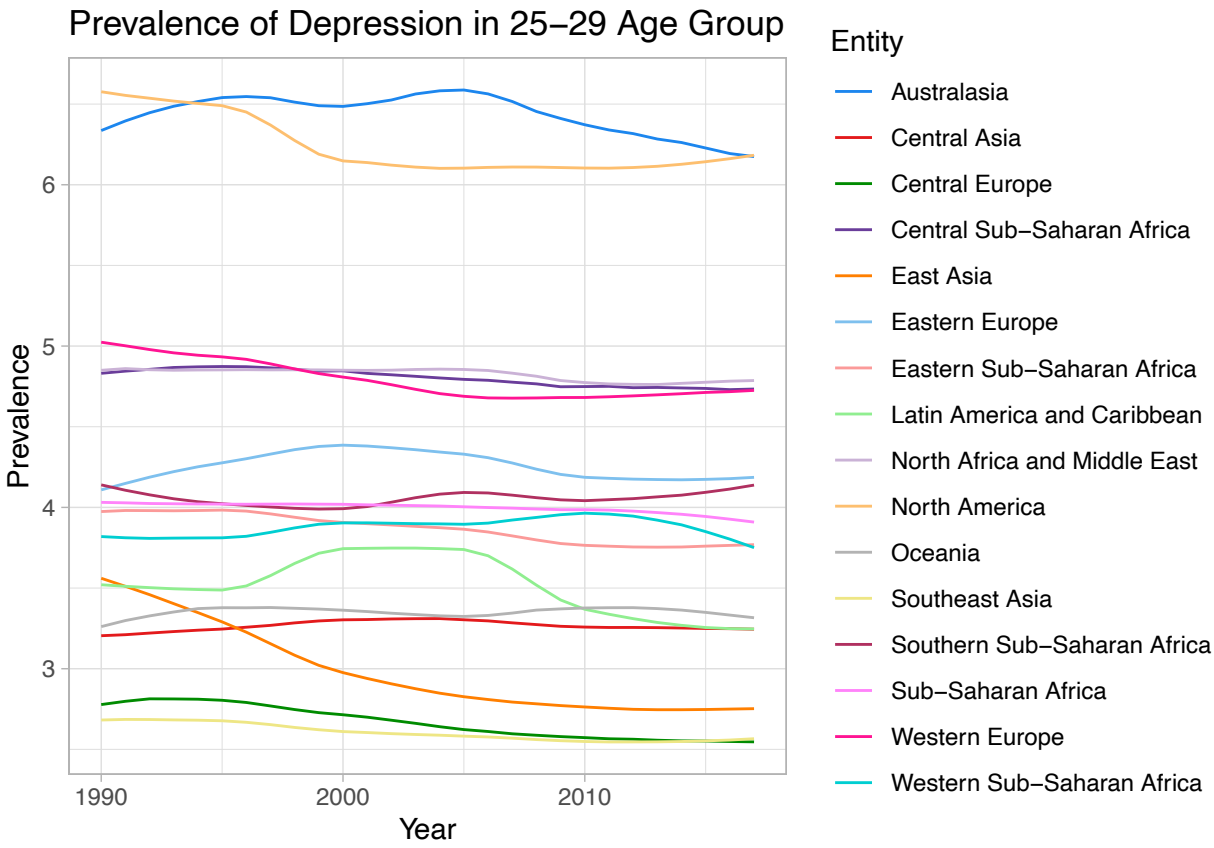
As I said before, we want to get a macro-level view of how prevalence is affecting different age groups across the world. Therefore, it would be useful to first build a line plot for each age group that compares the prevalence of depression in different regions.

I'm going to create a set of distinct colors for the line plots so that we can easily determine which country corresponds to each line. R's default color palettes often make it difficult to discern colors from each other, so creating our own distinct color palette will be useful. We'll need a palette with 16 distinct colors.

```
##create a set of colors to use for the regions in plots
region_cols <- c(c(
  "dodgerblue2",
  "#E31A1C", # red
  "green4",
  "#6A3D9A", # purple
  "#FF7F00", # orange
  "skyblue2",
  "#FB9A99", # lt pink
  "palegreen2",
  "#CAB2D6", # lt purple
  "#FDBF6F", # lt orange
  "gray70",
  "khaki2",
  "maroon",
  "orchid1",
  "deeppink1",
  "darkturquoise"))
```

Let's build an example plot for the 25-29 age group. We want to show prevalence on the y-axis, the Year on the x-axis, and each line should represent a different country. We should also be using the colors we jsut made for the lines.

```
##create example for 25-29 age group.
depression_age_regions %>%
  ggplot(aes(x=Year, y=twentyFive_to_twentyNine, group=Entity, color = Entity)) +
  geom_line() +
  labs(y = "Prevalence",
       title = "Prevalence of Depression in 25-29 Age Group") +
  theme_light() +
  scale_color_manual(values = region_cols)
```



This plot actually shows us a lot for the prevalence of depression in this age group. The first thing we notice is the big gap between the top two lines and the rest of the lines. The lines at the top appear to be Australasia and North America, so the prevalence of depression among 25-29 year-olds seems to be much higher in these two regions compared to the rest. I also noticed that the linear trend in prevalence over the time frame isn't as pronounced as I thought it would be. We see that East Asia shows a pretty significant and steady declining trend over time, and both the North American and Latin American/Caribbean regions show more bumps in their respective trends over time.

Let's make similar plots for the remainder of the age groups. Because we'll be making so many, we can create a function that will do this for us based on the age group we want to visualize. A couple things to note is that we'll need a title that changes based on the age group that we're looking at, and we're going to be using the depression_age_regions dataframe above within the function.

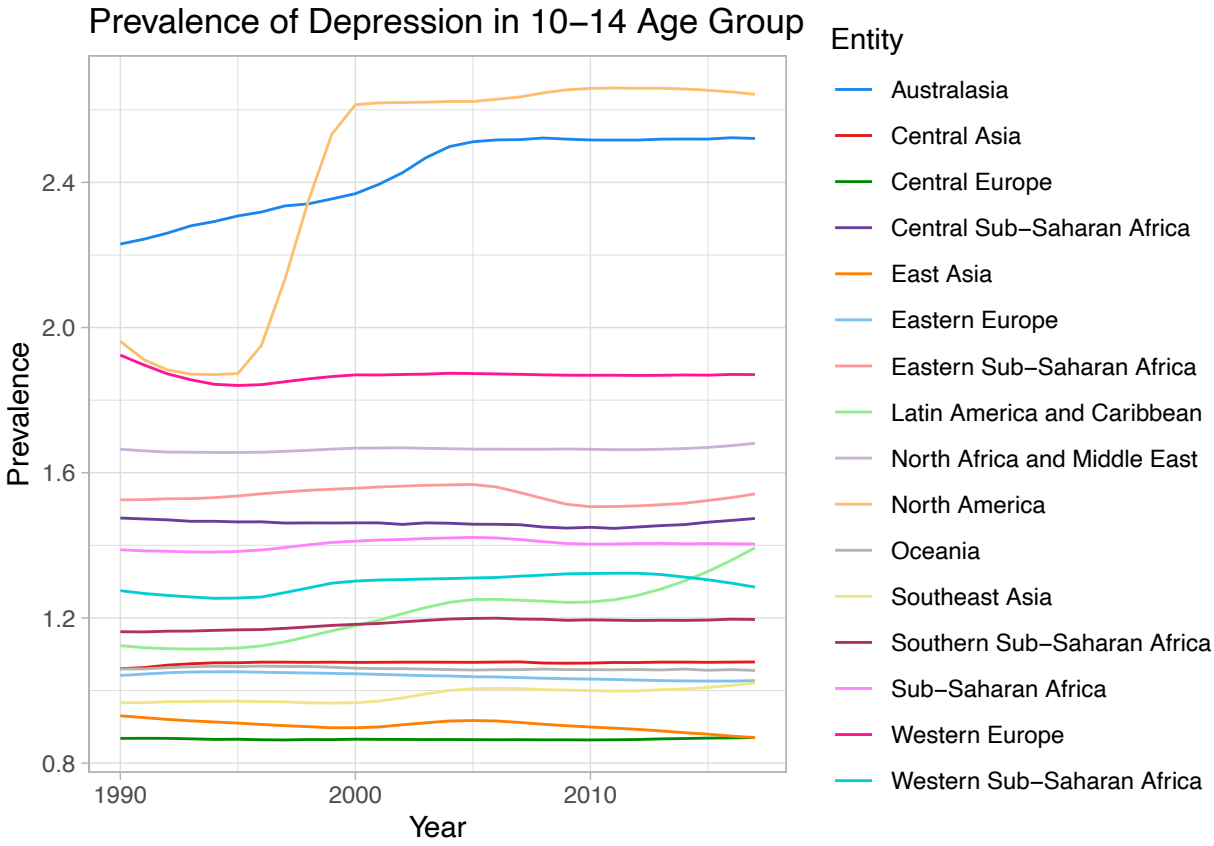
```
##create a custom function for building the graphs for prevalence in each region based on age group
build_age_graphs <- function(age_grp){

  ##determine which age group we're looking at and set age_title
  if(age_grp == depression_age_regions$ten_to_fourteen){
    age_title = "10-14"
  }
  if(age_grp == depression_age_regions$fifteen_to_nineteen){
    age_title = "15-19"
  }
  if(age_grp == depression_age_regions$twenty_to_twentyFour){
    age_title = "20-24"
  }
  if(age_grp == depression_age_regions$twentyFive_to_twentyNine){
    age_title = "25-29"
```

```
  }
  if(age_grp == depression_age_regions$thirty_to_thirtyFour){
    age_title = "30-34"
  }
  if(age_grp == depression_age_regions$fifty_to_sixtyNine){
    age_title = "50-69"
  }
  if(age_grp == depression_age_regions$seventy_plus){
    age_title = "70+"
  }

  ##build the graph with the user-defined age_grp as the values to plot for prevalence
  ##we still want Year on x-axis and Entity to be group and color
  depression_age_regions %>%
    ggplot(aes(x=Year, y=age_grp, group=Entity, color = Entity)) +
    geom_line(size=.5) +
    labs(y = "Prevalence",
       title = paste("Prevalence of Depression in", age_title, "Age Group")) +
  theme_light() +
  scale_color_manual(values = region_cols)
}
```

Let's build a graph for each of the age groups starting with 10-14.

```
##10-14 grap
build_age_graphs(depression_age_regions$ten_to_fourteen)
```

```
## Warning in if (age_grp == depression_age_regions$ten_to_fourteen) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$fifteen_to_nineteen) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$twenty_to_twentyFour) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$twentyFive_to_twentyNine) {:
## the condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$thirty_to_thirtyFour) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$fifty_to_sixtyNine) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$seventy_plus) {: the condition
## has length > 1 and only the first element will be used
```



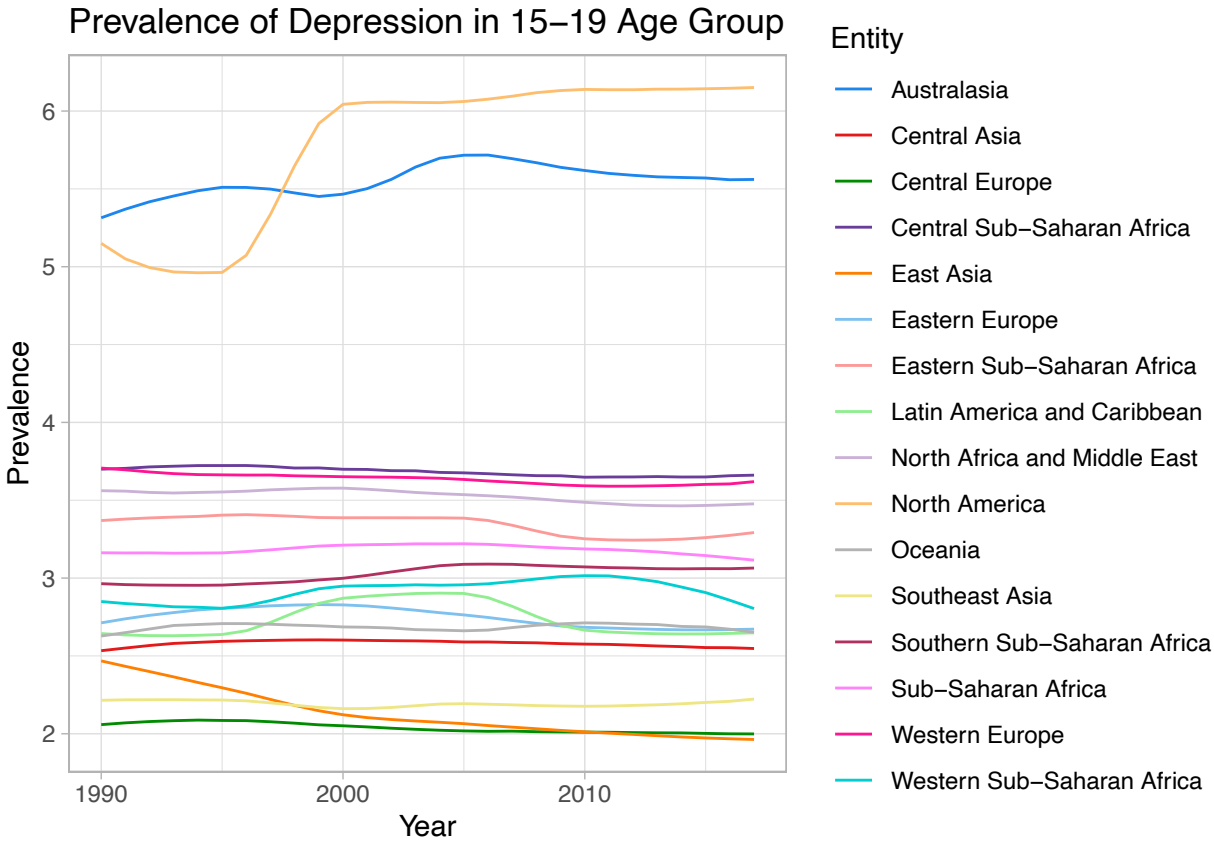
We can make some similar conclusions about 10-14 year-olds as we did with the 25-29 age group. The first thing we notice is the huge increase in prevalence of depression over the 1995-2000 time frame in North America, and Australasia also shows a steady incline from 1990-2005. These two regions also have much higher prevalence compared to other regions for this age group.

We'll build the graph for the 15-19 age group.

```
##plot for 15-19 year age group
build_age_graphs(depression_age_regions$fifteen_to_nineteen)
```

```
## Warning in if (age_grp == depression_age_regions$ten_to_fourteen) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$fifteen_to_nineteen) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$twenty_to_twentyFour) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$twentyFive_to_twentyNine) {:
## the condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$thirty_to_thirtyFour) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$fifty_to_sixtyNine) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$seventy_plus) {: the condition
## has length > 1 and only the first element will be used
```



Again, we see a similar longitudinal trend in North America as we saw for the 10-14 age group. We also notice that North America and Australasia are set far apart from other regions for 15-19 year olds.

Let's now look at the 20-24 age group.

```
##plot for 20-24 year age group
build_age_graphs(depression_age_regions$twenty_to_twentyFour)
```

```
## Warning in if (age_grp == depression_age_regions$ten_to_fourteen) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$fifteen_to_nineteen) {: the
## condition has length > 1 and only the first element will be used
```

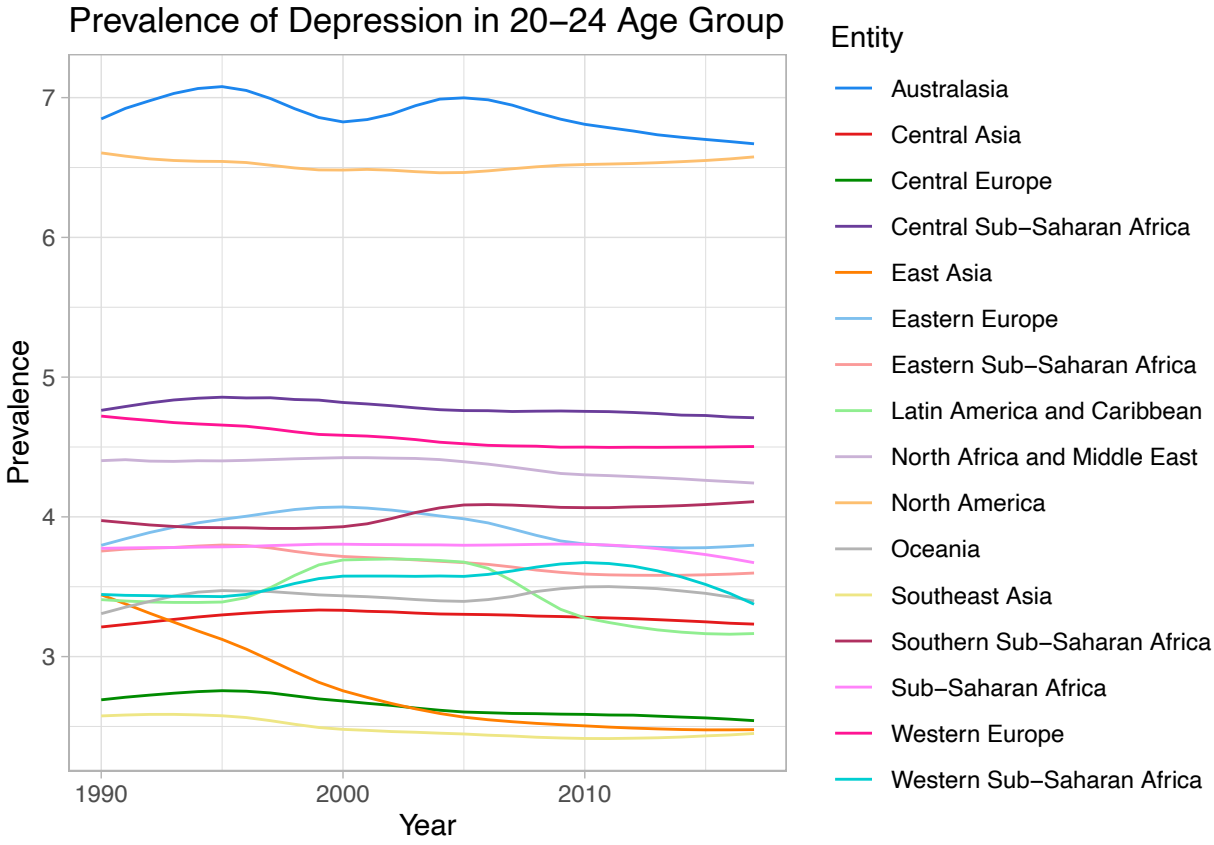
```
## Warning in if (age_grp == depression_age_regions$twenty_to_twentyFour) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$twentyFive_to_twentyNine) {:
## the condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$thirty_to_thirtyFour) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$fifty_to_sixtyNine) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$seventy_plus) {: the condition
## has length > 1 and only the first element will be used
```



Like we saw with the 25-29 year-olds, East Asia shows a steady decline in prevalence over time. And, as we've seen in the previous age groups, North America and Australasia show much highe prevalence compared to other age groups.

Since we made the graph for 25-29 age group, let's skip to the 30-34 age group now.

```
##plot for 30-34 year age group
build_age_graphs(depression_age_regions$thirty_to_thirtyFour)
```

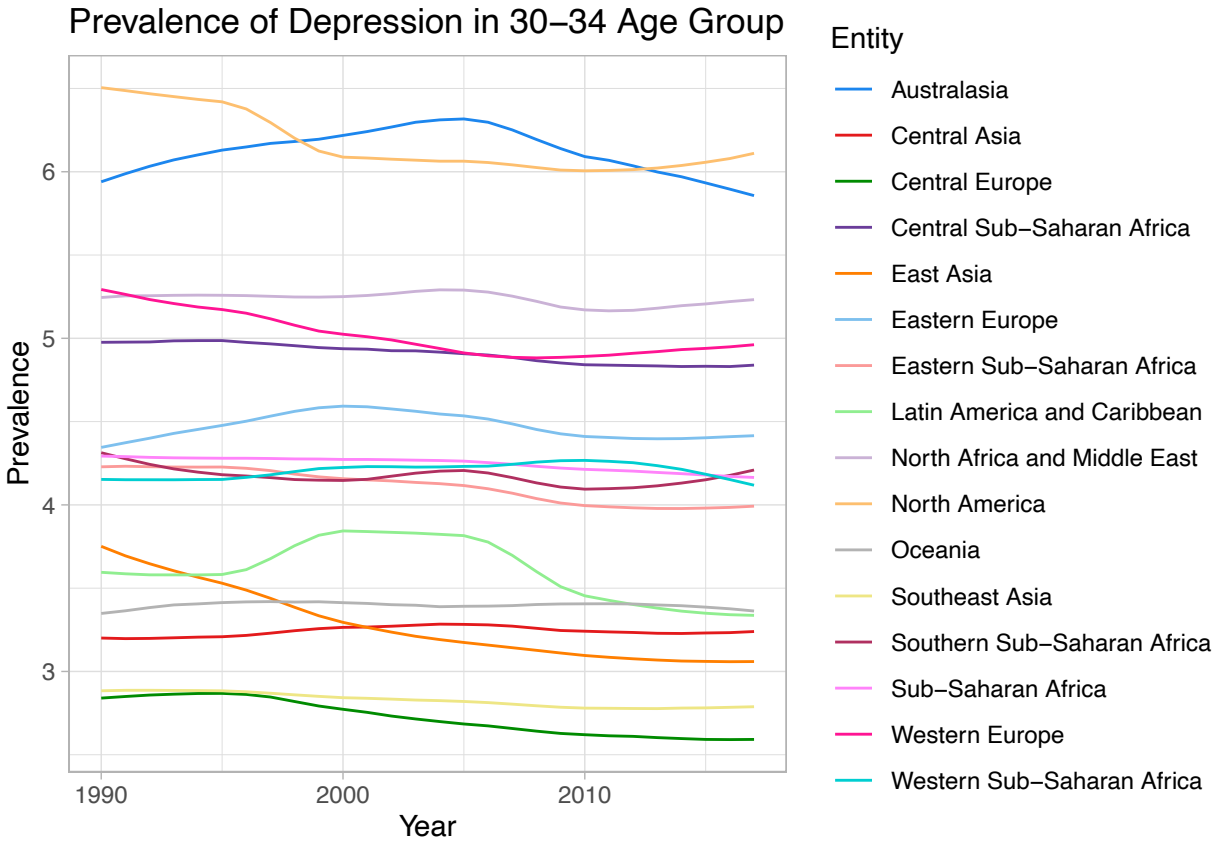
```
## Warning in if (age_grp == depression_age_regions$ten_to_fourteen) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$fifteen_to_nineteen) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$twenty_to_twentyFour) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$twentyFive_to_twentyNine) {:
## the condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$thirty_to_thirtyFour) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$fifty_to_sixtyNine) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$seventy_plus) {: the condition
## has length > 1 and only the first element will be used
```



Although the gap is closing, we again see that North America and Australasia have a higher prevalence compared to other regions. The longitudinal trends are similar to to those we've seen before, but we see Australasia peaking in 2005 and ten declining through 2019, and North America has decreasing prevalence with a small increase around 2016-2017.

Now let's look at the 50-69 and 70+ age groups.

```
##plot for 50-69 year age group
build_age_graphs(depression_age_regions$fifty_to_sixtyNine)
```

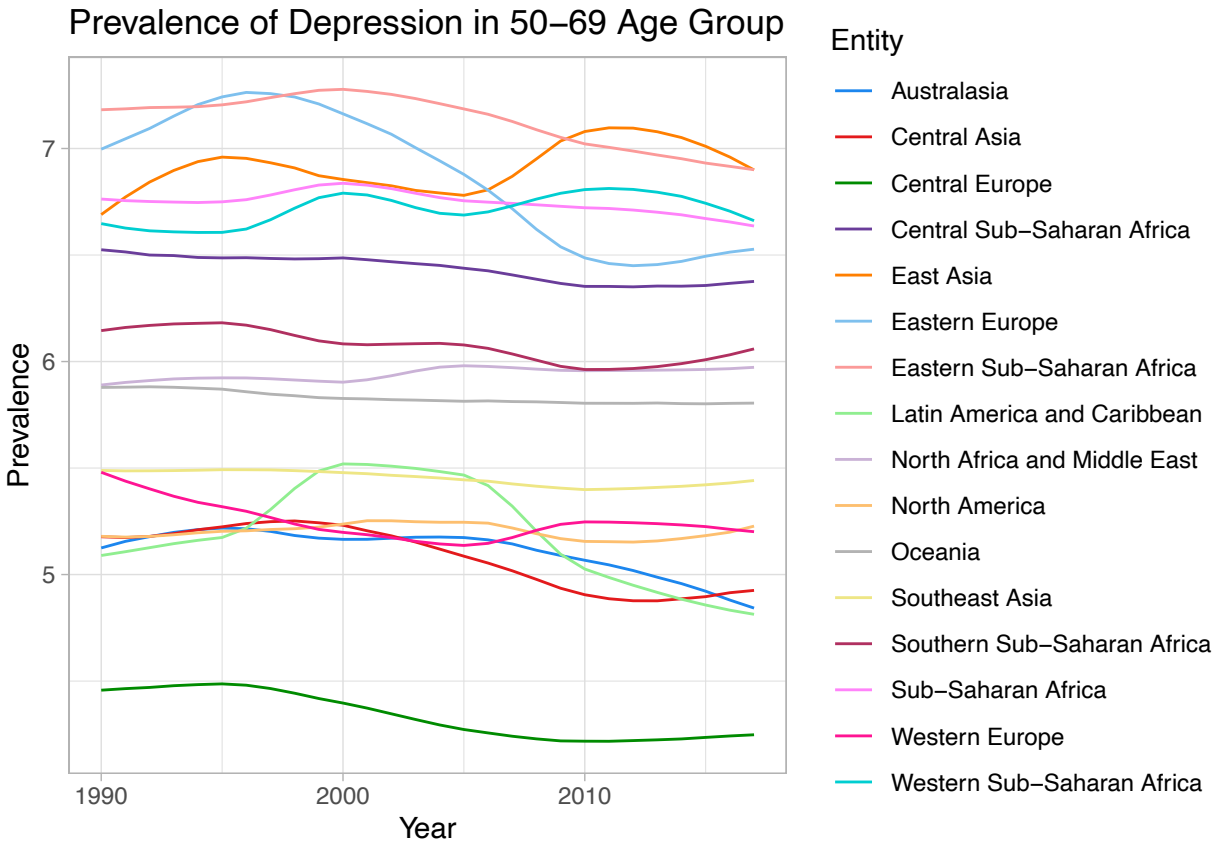
```
## Warning in if (age_grp == depression_age_regions$ten_to_fourteen) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$fifteen_to_nineteen) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$twenty_to_twentyFour) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$twentyFive_to_twentyNine) {:
## the condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$thirty_to_thirtyFour) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$fifty_to_sixtyNine) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$seventy_plus) {: the condition
## has length > 1 and only the first element will be used
```



The picture is starting to change with this age group- North America and Australasia are no longer set apart from the rest, and we see other regions with a much higher prevalence compared to the others.

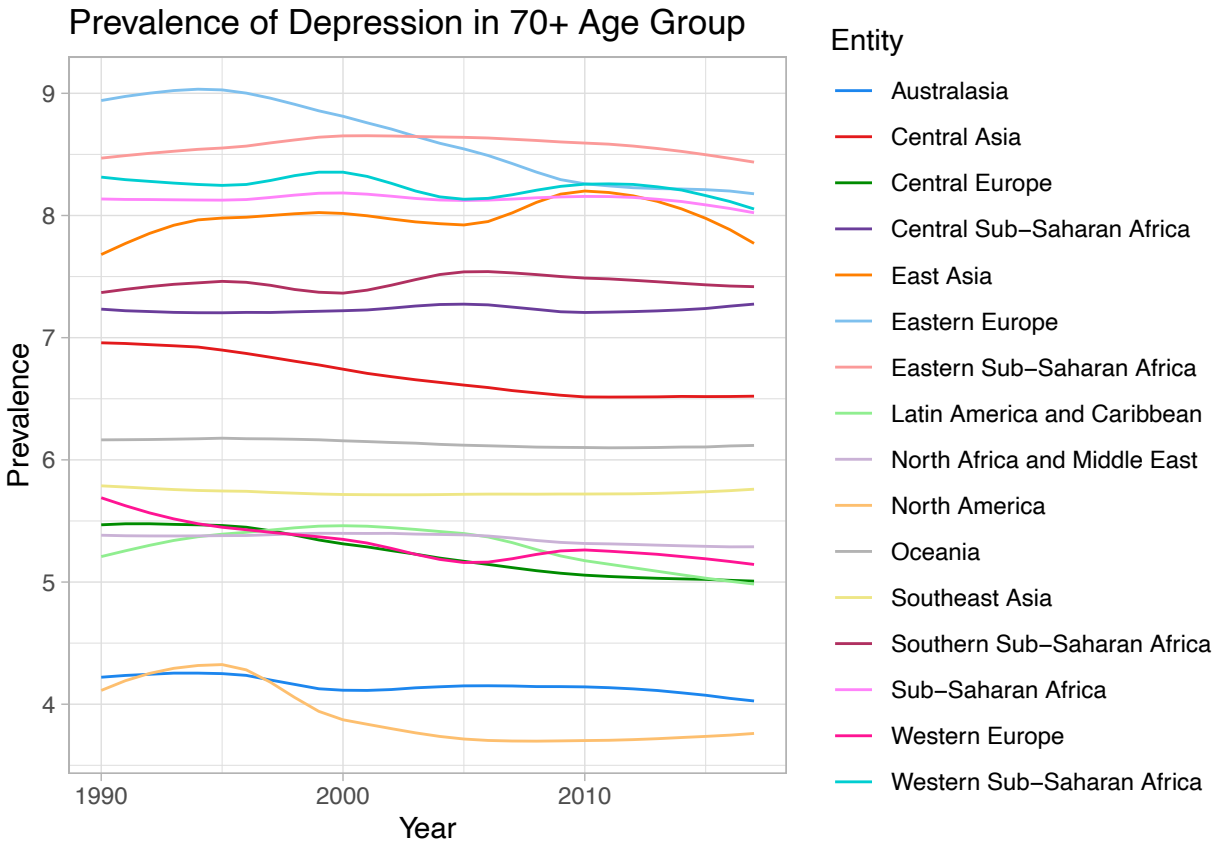
Let's see if we see something similar for the 70+ age group.

```
##plot for 70+ year age group
build_age_graphs(depression_age_regions$seventy_plus)
```

```
## Warning in if (age_grp == depression_age_regions$ten_to_fourteen) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$fifteen_to_nineteen) {: the
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (age_grp == depression_age_regions$twenty_to_twentyFour) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$twentyFive_to_twentyNine) {:
## the condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$thirty_to_thirtyFour) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$fifty_to_sixtyNine) {: the
## condition has length > 1 and only the first element will be used

## Warning in if (age_grp == depression_age_regions$seventy_plus) {: the condition
## has length > 1 and only the first element will be used
```



So now Australasia and North America actually have the lowest prevalence of depression for the 70+ age group. The picture has flipped where the African and some Asian and European regions have the highest prevalence of depression for the older age groups.

We'll make a table that summarizes the prevalence of depression over age groups in each country. We're going to use the average prevalence over the 1990-2017 time frame. Let's group by Entity and the summarize by finding the average prevalence for each age group.

```
##make table with average prevalence over the timeframe for each age group in all regions
depression_age_regions_table <- depression_age_regions %>%
  group_by(Entity) %>%
```

```
  summarise(AvgTenFourteen = mean(ten_to_fourteen),
            AvgFifteenNineteen = mean(ten_to_fourteen),
            AvgTwentyTwentyFour = mean(twenty_to_twentyFour),
            AvgTwentyfiveTwentynine = mean(twentyFive_to_twentyNine),
            AvgThirtyThirtyfour = mean(thirty_to_thirtyFour),
            AvgFiftySixtynine = mean(fifty_to_sixtyNine),
            AvgSeventyPlus = mean(seventy_plus))

depression_age_regions_table
```

```
## # A tibble: 16 x 8
##    Entity     AvgTenFourteen AvgFifteenNinet~ AvgTwentyTwenty~ AvgTwentyfiveTwe~
##    <fct>               <dbl>            <dbl>            <dbl>              <dbl>
##  1 Australas~          2.42             2.42              6.88               6.43
##  2 Central A~          1.08             1.08              3.28               3.27
##  3 Central E~          0.866            0.866             2.64               2.67
##  4 Central S~          1.46             1.46              4.78               4.80
##  5 East Asia           0.903            0.903             2.76               2.99
##  6 Eastern E~          1.04             1.04              3.92               4.26
##  7 Eastern S~          1.54             1.54              3.68               3.87
##  8 Latin Ame~          1.21             1.21              3.44               3.52
##  9 North Afr~          1.66             1.66              4.36               4.83
## 10 North Ame~          2.42             2.42              6.52               6.24
## 11 Oceania             1.06             1.06              3.44               3.35
## 12 Southeast~          0.988            0.988             2.48               2.60
## 13 Southern ~          1.18             1.18              4.01               4.06
## 14 Sub-Sahar~          1.40             1.40              3.78               4.00
## 15 Western E~          1.87             1.87              4.57               4.79
## 16 Western S~          1.30             1.30              3.54               3.88
## # ... with 3 more variables: AvgThirtyThirtyfour <dbl>,
## #   AvgFiftySixtynine <dbl>, AvgSeventyPlus <dbl>
```

This table helps confirm our observations from the graphs we made above. While North America and Australasia show higher prevalences in younger and middle-aged groups, Africa and a couple Europe and Asia regions show a much higher prevalence for the older age groups.

Sub-Analysis: North America and Australasia

Let's do a sub-analysis for North America and Australasia. We're going to create line plots of the prevalence for each age group in these two regions to have a more direct view of the comparisons between ages. These plots will look very similar to the plots we made above, but instead of the groups being entities, the groups will be the age groups, and we're going to look at one region (entity) at a time.

Because of the way that our data is currently set up, we need to pivot the data for a single region. This puts the year, age group, and prevalence all in a row rather than having the year and specific age groups as columns. Our data should have a column for year, a column with the age group, and a column with the prevalence. We'll make a function to do this since we have to do it several times in analyses later on. The function should take the entity as a user-defined parameter so that we're looking at a single region. We can also change the names of the age groups to something like 10-14 so that when we make plots the labels on the axes use raw numbers rather than whatever the column names were before pivoting.

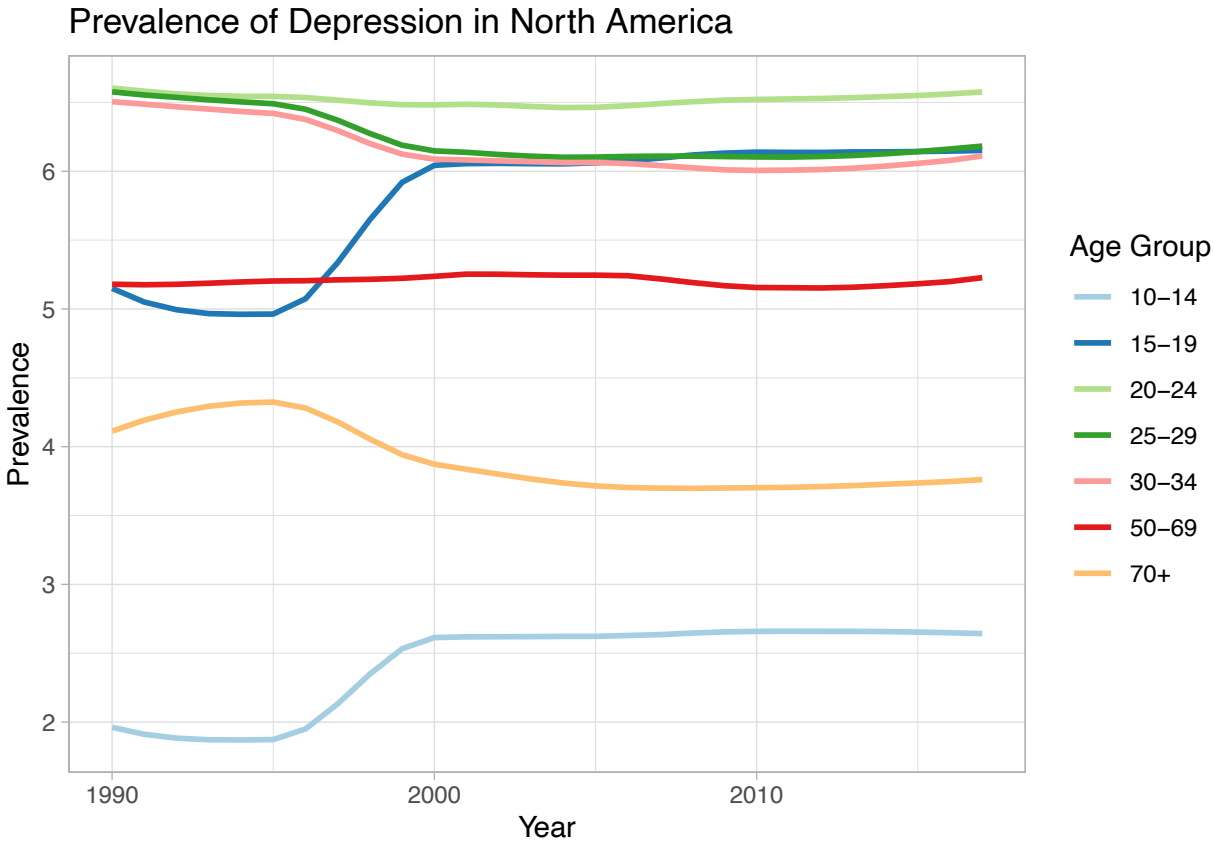
```
##create a function for pivoting the data
##take entity as a parameter
pivot_my_data <- function(entity){
  depression_age_regions %>%
  filter(Entity == entity) %>% ##filter for just the region we're looking at
  select(Year, ten_to_fourteen, fifteen_to_nineteen, twenty_to_twentyFour, twentyFive_to_twentyNine, th
    pivot_longer(!Year, names_to = "age_group", values_to = "prevalence") %>% ##pivot so age_group is a
    mutate(age_group = recode(age_group, "fifteen_to_nineteen" = "15-19", ##change the names of the age
                                "ten_to_fourteen" = "10-14",
                                 "twenty_to_twentyFour" = "20-24",
                                "twentyFive_to_twentyNine" = "25-29",
                                "thirty_to_thirtyFour" = "30-34",
                                "fifty_to_sixtyNine" = "50-69",
                                "seventy_plus" = "70+"))
}
```

Now we can pivot the data for North America.

```
##pivot the data for North America
depression_northAmerica_pivot <- pivot_my_data("North America")
```

We can use the pivoted dataframe to build another line graph. This will show the prevalence of depression by age group over the time period in North America.

```
##plot comparing North America age group prevalences
depression_northAmerica_pivot %>%
  ggplot(aes(x=Year, y=prevalence, group=age_group, color = age_group)) +
  geom_line(size = 1) +
  labs(y = "Prevalence",
       title = "Prevalence of Depression in North America",
       color = "Age Group") +
  scale_colour_brewer(palette = "Paired") +
  theme_light()
```



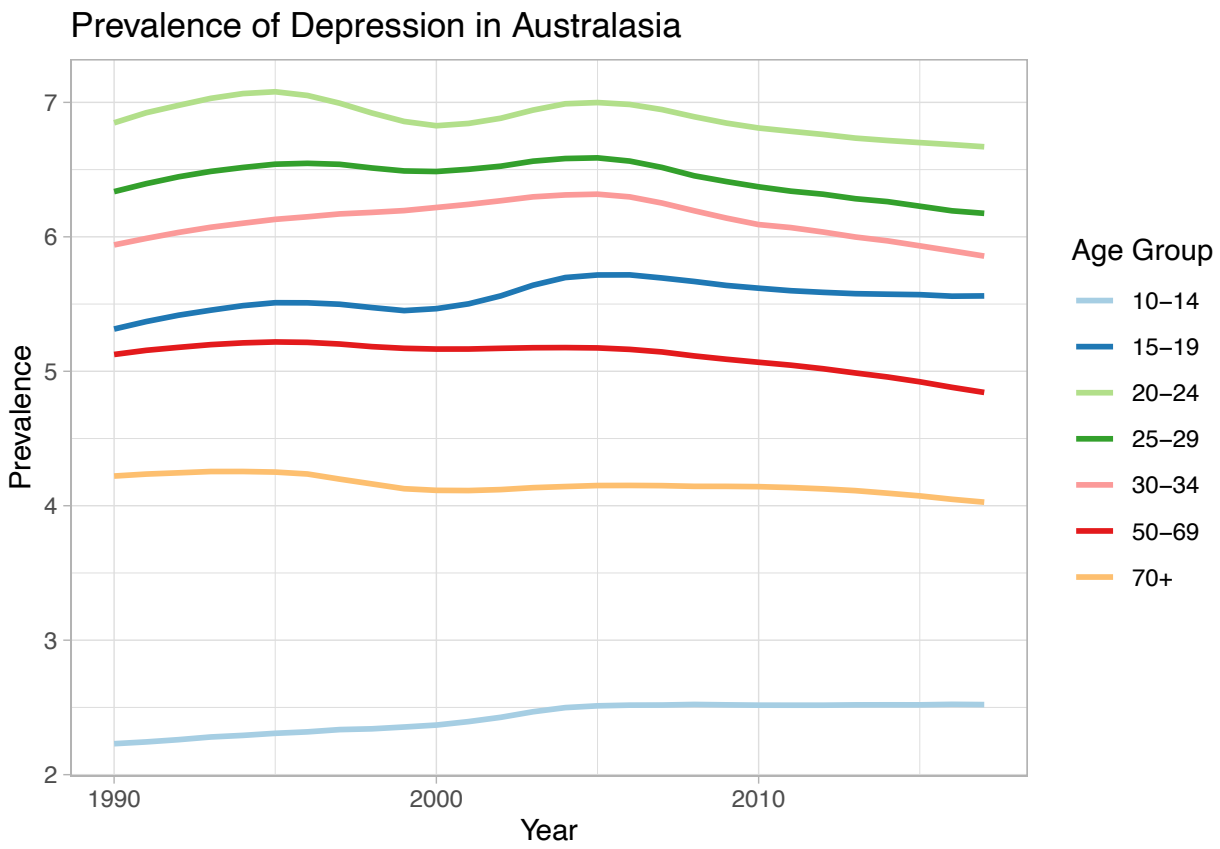
From this graph, we see that in North America, the 20-24, 25-29, 30-34, and 15-19 age groups have the highest prevalences in the region. We also notice that, right around 1995, the prevalence of depression among the 10-14 age groups and 15-19 age groups starts increasing, and it peaks in about 2000. The 70+ age group also has a lower prevalence of depression compared to these other age groups, and we saw that it's much lower compared to other regions.

Before we move on to Australasia, let's turn the code for that North America graph into a function since we're going to continue making them for different regions. We'll need to pass in both an entity name and the pivoted dataframe for that entity as parameters of the function.

```
##create a function that will create the line graph for prevalence in each age group for a specific reg
graph_for_region <- function(entity, df){
  df %>%
    ggplot(aes(x=Year, y=prevalence, group=age_group, color = age_group)) +
    geom_line(size =1) +
    labs(y = "Prevalence",
         title = paste("Prevalence of Depression in", entity),
         color = "Age Group") +
    scale_colour_brewer(palette = "Paired") +
    theme_light()
}
```

Australasia also had much higher prevalences among the younger and middle-aged groups compared to other regions, so I'm going to make a similar graph to see how prevalence of depression is changing over time within this region. We need to pivot our data the same way we pivoted the data for North America, so we can call the pivot_my_data function, and then make a graph with the function we just created.

```
##pivot the data for Australasia
depression_australasia_pivot <- pivot_my_data("Australasia")

graph_for_region("Australasia", depression_australasia_pivot)
```



Within Australasia, the 20-24, 25-29, 30-34, 15-19, and 50-69 age groups have a much higher prevalence of depression compared to the 70+ and 10-14 age groups, which is similar to the pattern in North America.

I'll build a couple of regression models to determine whether age group plays a role in prevalence of depression in these two regions. One model will predict prevalence of depression in North America using age group as the only covariate, and the second model will predict prevalence of depression in Australasia using age group as the only covariate. Since the 70+ age group has the second lowest prevalence in the graph above, I'm going to make them the reference group in order to determine how much higher prevalence is among the adolescent and middle-aged groups compared to the oldest age group. In order to force a certain reference group, we can first set the levels for the age groups so that 70+ comes first and goes in descending age group order from there, and we'll do this in the pivoted dataframe for North America.

```
##relevel the factors to make 70+ the oldest age group and then go in decreasing order by age group
depression_northAmerica_pivot$age_group <- factor(depression_northAmerica_pivot$age_group,
                                                  levels=c("70+",
                                                           "50-69",
                                                           "30-34",
                                                           "25-29",
                                                           "20-24",
                                                           "15-19",
                                                           "10-14"))
##make sure age_group is still a factor
```

```
depression_northAmerica_pivot$age_group <- as.factor(depression_northAmerica_pivot$age_group)
##make linear model predicting prevalence with age group in North America
north_america_mod <- lm(prevalence~age_group, data = depression_northAmerica_pivot)
summary(north_america_mod)
```

```
##
## Call:
## lm(formula = prevalence ~ age_group, data = depression_northAmerica_pivot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82253 -0.12599 -0.00575  0.21052  0.42130
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.90299    0.04826   80.88   <2e-16 ***
## age_group50-69  1.29973    0.06824   19.05   <2e-16 ***
## age_group30-34  2.26389    0.06824   33.17   <2e-16 ***
## age_group25-29  2.33485    0.06824   34.21   <2e-16 ***
## age_group20-24  2.61812    0.06824   38.36   <2e-16 ***
## age_group15-19  1.88078    0.06824   27.56   <2e-16 ***
## age_group10-14 -1.47936    0.06824  -21.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2553 on 189 degrees of freedom
## Multiple R-squared:  0.9684, Adjusted R-squared:  0.9674
## F-statistic: 965.2 on 6 and 189 DF,  p-value: < 2.2e-16
```

The prevalence of depression increases by at least 2% in middle-aged groups compared to the 70+ age group holding all else constant. The coefficients are also all statistically significant, so there are meaningful differences between prevalence of depression for these age groups.

This link showed me the code to change the order of the age group factors to something meaninfgul in the context of what we want to show: https://stats.stackexchange.com/questions/430770/in-a-multilevel-linear-regression-how-does-the-reference-level-affect-other-lev.

Let's make a similar model for Australasia and relevel the groups in the pivoted dataframe for Australasia as we did before for North America. Again, age group is going to predict prevalence.

```
##relevel groups so 70+ is reference group
depression_australasia_pivot$age_group <- factor(depression_australasia_pivot$age_group,
                                                 levels=c("70+",
                                                          "50-69",
                                                          "30-34",
                                                          "25-29",
                                                          "20-24",
                                                          "15-19",
                                                          "10-14"))
##build linear model to predict prevalence with age group
australasia_mod <- lm(prevalence~age_group, data = depression_australasia_pivot)
summary(australasia_mod)
```

```
##
```

```
## Call:
## lm(formula = prevalence ~ age_group, data = depression_australasia_pivot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26866 -0.07567  0.01120  0.09087  0.19772
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.15368    0.02082  199.50   <2e-16 ***
## age_group50-69  0.95765    0.02944   32.52   <2e-16 ***
## age_group30-34  1.96573    0.02944   66.76   <2e-16 ***
## age_group25-29  2.28089    0.02944   77.47   <2e-16 ***
## age_group20-24  2.73063    0.02944   92.74   <2e-16 ***
## age_group15-19  1.39737    0.02944   47.46   <2e-16 ***
## age_group10-14 -1.73020    0.02944  -58.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1102 on 189 degrees of freedom
## Multiple R-squared:  0.9942, Adjusted R-squared:  0.9941
## F-statistic:  5433 on 6 and 189 DF,  p-value: < 2.2e-16
```

We see a similar result with this model where the middle-aged groups have a 2% higher prevalence compared to the 70+ age group. All of the coefficients are also significant, so we know that there are statistically different prevalences between these age groups.

Subanalysis: African Regions

Earlier we noted that the regions with the highest prevalence of depression switches when we look at some of the older age groups. Specifically, within the age 50-69 and 70+ age groups, some of the African, European, and Asian regions show much higher prevalence compared to the other regions. In fact, North America and Australasia have the two lowest prevalences in the 70+ age group and some of the lowest in the 50-69 age group.

Let's do a subanalysis of the regions with the highest prevalences in the 50-69 age group and the 70+ age group. We'll first determine which regions in these two age groups have an average prevalence of depression over the time frame above 6% in the 50-69 age group and above 7% in the 70+ age group. In the 50-69 age group the highest prevalences are between 6-8%, so we want to make sure that we capture enough of the regions with the higher prevalences. Similarly, in the 70+ age group, the highest prevalences fall in the 7-9% range, so we want to make sure we're looking at all regions with prevalences greater than 7%.

We'll make two tables, one for the 50-69 age group and one for the 70+ age group. In the 50-69 age group table, we just want regions with prevalence higher than 6%, and in the 70+ age group table, we just want regions with prevalence higher than 7%.

```
##which regions have an average prevalence greater than 7% over the 1990-2017 time frame in the 50-59 a

##subset the depression_age_regions_table to include just the 50-69 column
##filter to include only regions with a prevalence greater than 6%
depression_age_regions_table %>%
  select(Entity, AvgFiftySixtynine) %>%
  filter(AvgFiftySixtynine > 6)
```

```
## # A tibble: 7 x 2
##   Entity                       AvgFiftySixtynine
##   <fct>                                    <dbl>
## 1 Central Sub-Saharan Africa               6.43
## 2 East Asia                                6.91
## 3 Eastern Europe                           6.87
## 4 Eastern Sub-Saharan Africa               7.13
## 5 Southern Sub-Saharan Africa              6.07
## 6 Sub-Saharan Africa                       6.75
## 7 Western Sub-Saharan Africa               6.71
```

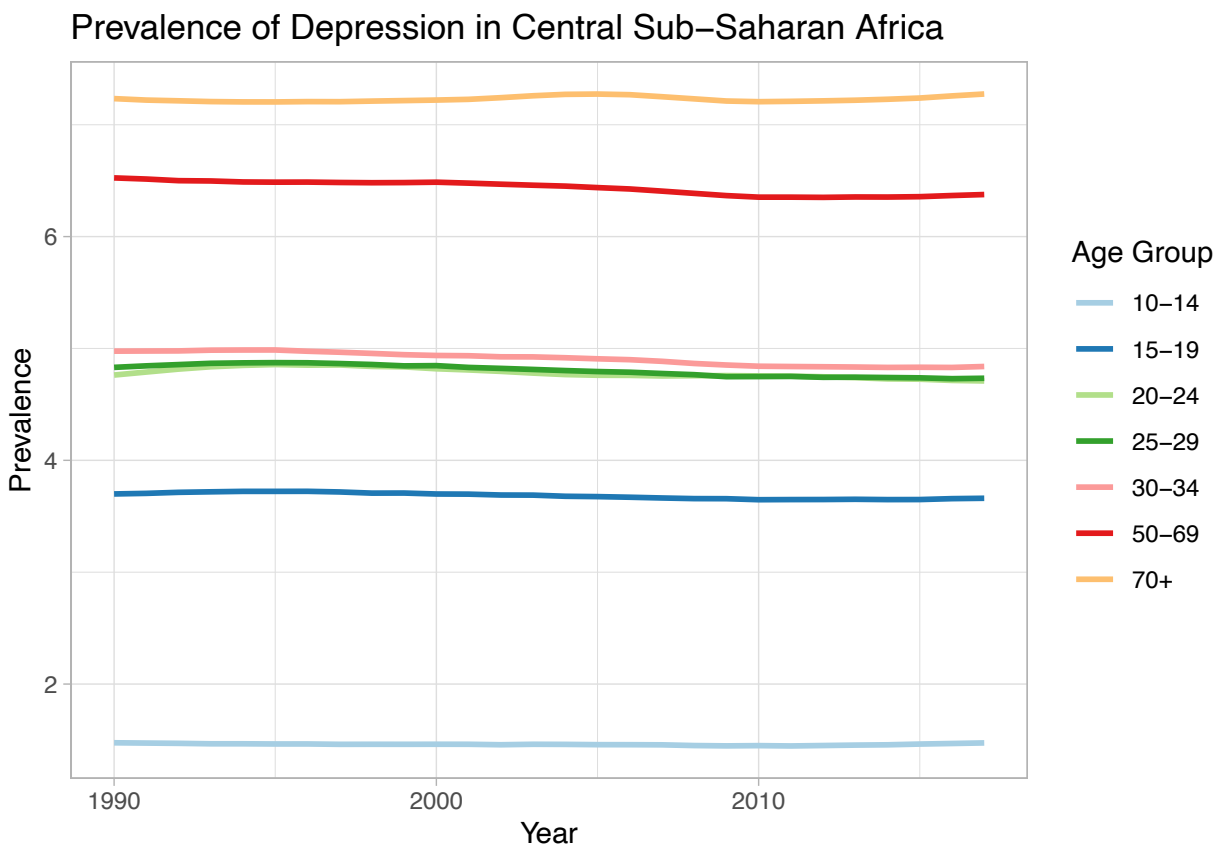
```
##subset the depression_age_regions_table to include just the 70+ column
##filter to include only regions with a prevalence greater than 7%
depression_age_regions_table %>%
  select(Entity, AvgSeventyPlus) %>%
  filter(AvgSeventyPlus > 7)
```

```
## # A tibble: 7 x 2
##   Entity                       AvgSeventyPlus
##   <fct>                                 <dbl>
## 1 Central Sub-Saharan Africa            7.23
## 2 East Asia                             7.99
## 3 Eastern Europe                        8.61
## 4 Eastern Sub-Saharan Africa            8.57
## 5 Southern Sub-Saharan Africa           7.45
## 6 Sub-Saharan Africa                    8.13
## 7 Western Sub-Saharan Africa            8.24
```

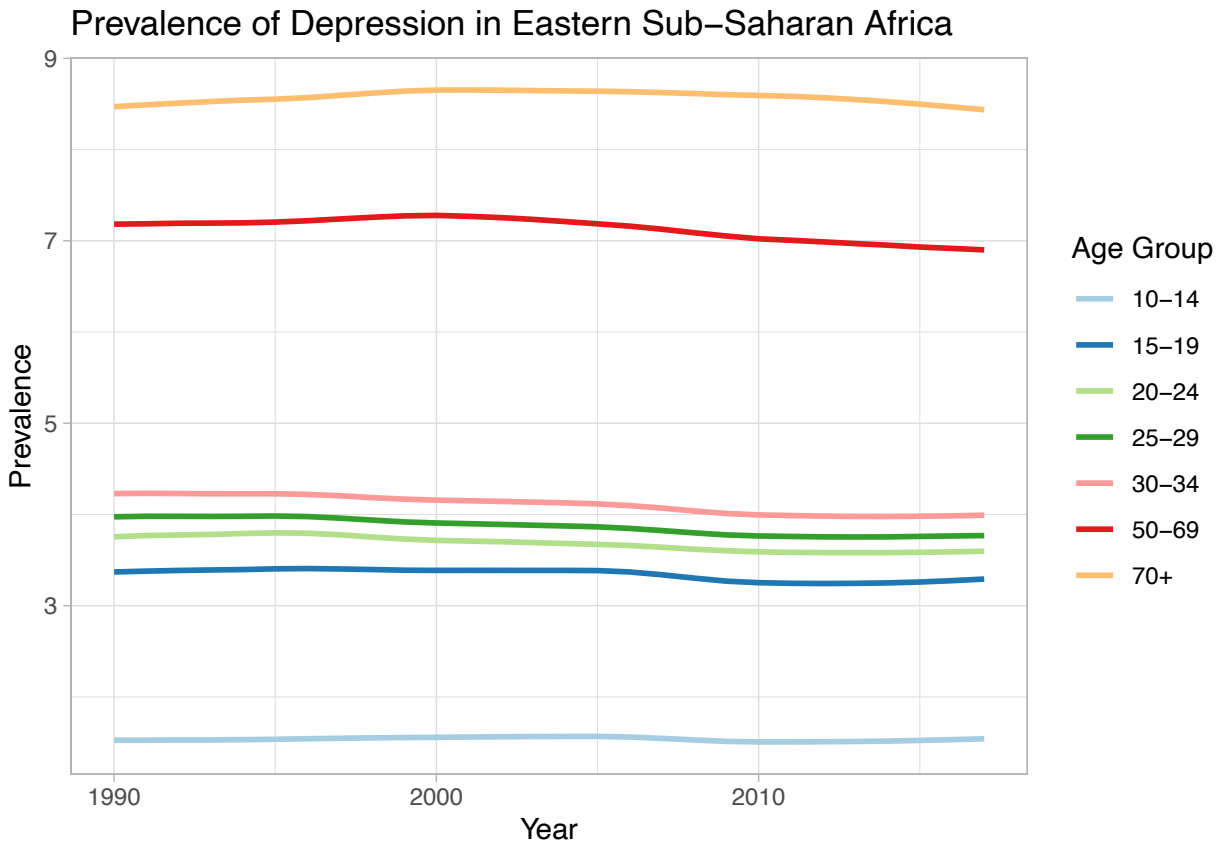
Based on these two tables above, it looks like the African and some of the European and Asian regions have the highest prevalence of depression in these two older age groups. Let's do a subanalysis of just the African regions similar to the one we did for Australasia and North America.

We'll first make graphs for each of the African regions in the tables above to see how prevalence is changing over time for each of the age groups. We'll need to pivot the data like we did for Australasia and North America, so we can call the function for doing so on each of the African regions then use the graph function for making the graphs of individual regions.

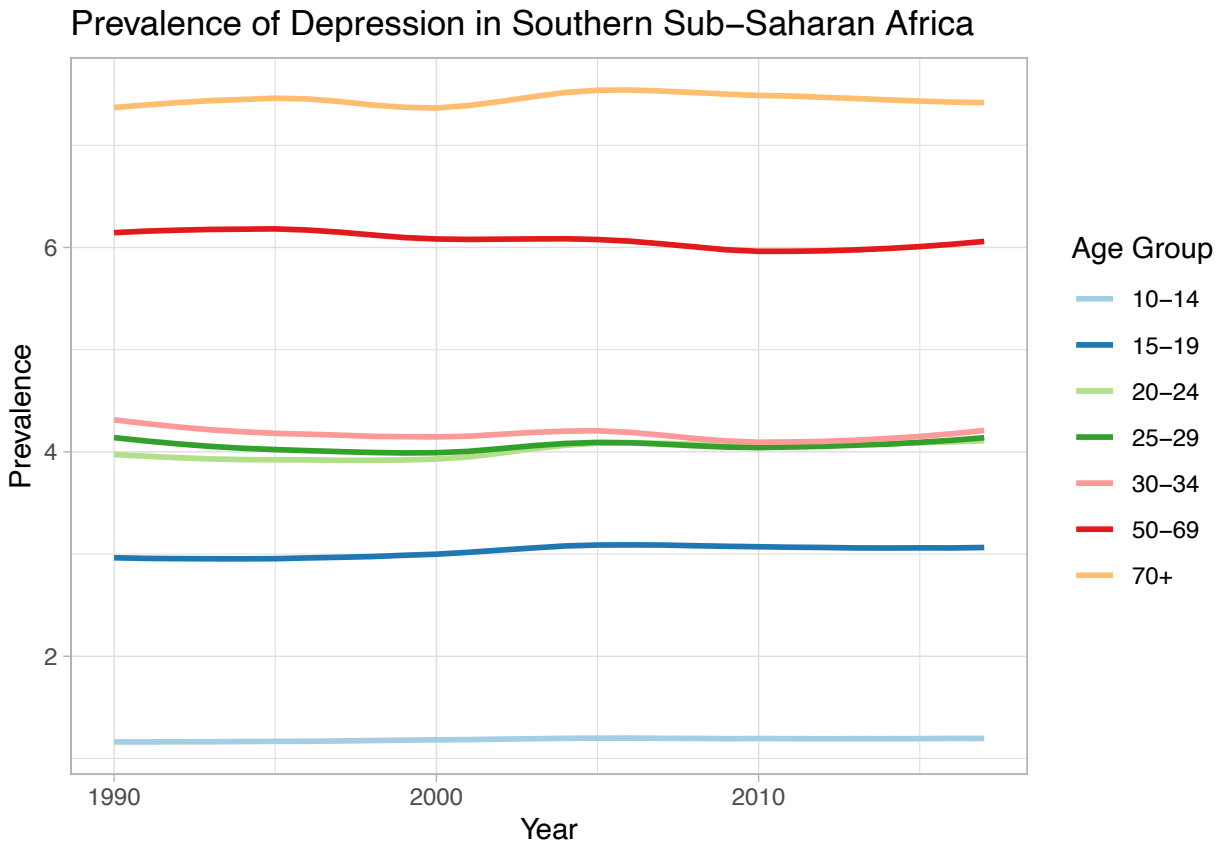
```
##pivot the data for Central Sub-Saharan Africa
central_subSaharan_pivot <- pivot_my_data("Central Sub-Saharan Africa")

##make the graph for pivoted data
graph_for_region("Central Sub-Saharan Africa", central_subSaharan_pivot)
```



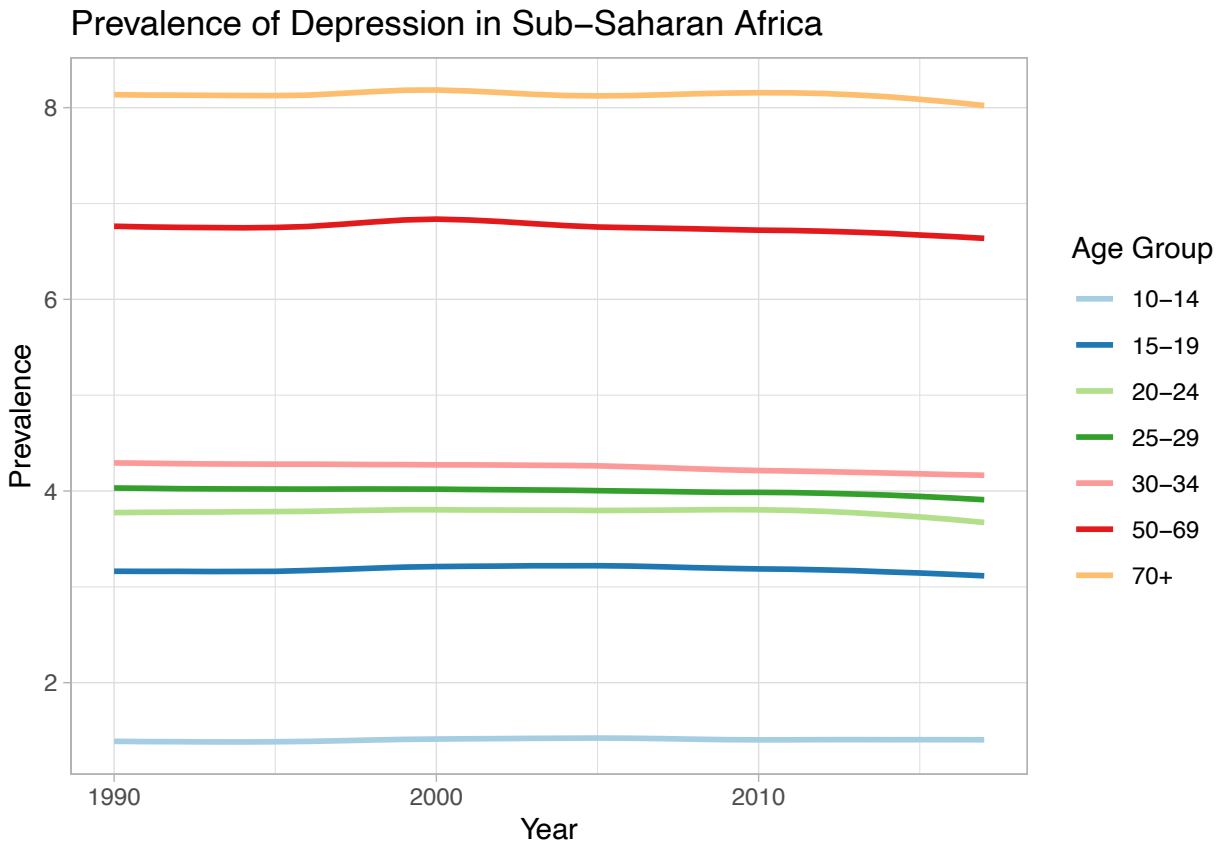

```
##pivot the data for Eastern Sub-Saharan Africa
east_subSaharan_pivot <- pivot_my_data("Eastern Sub-Saharan Africa")

##make the graph for pivoted data
graph_for_region("Eastern Sub-Saharan Africa", east_subSaharan_pivot)
```



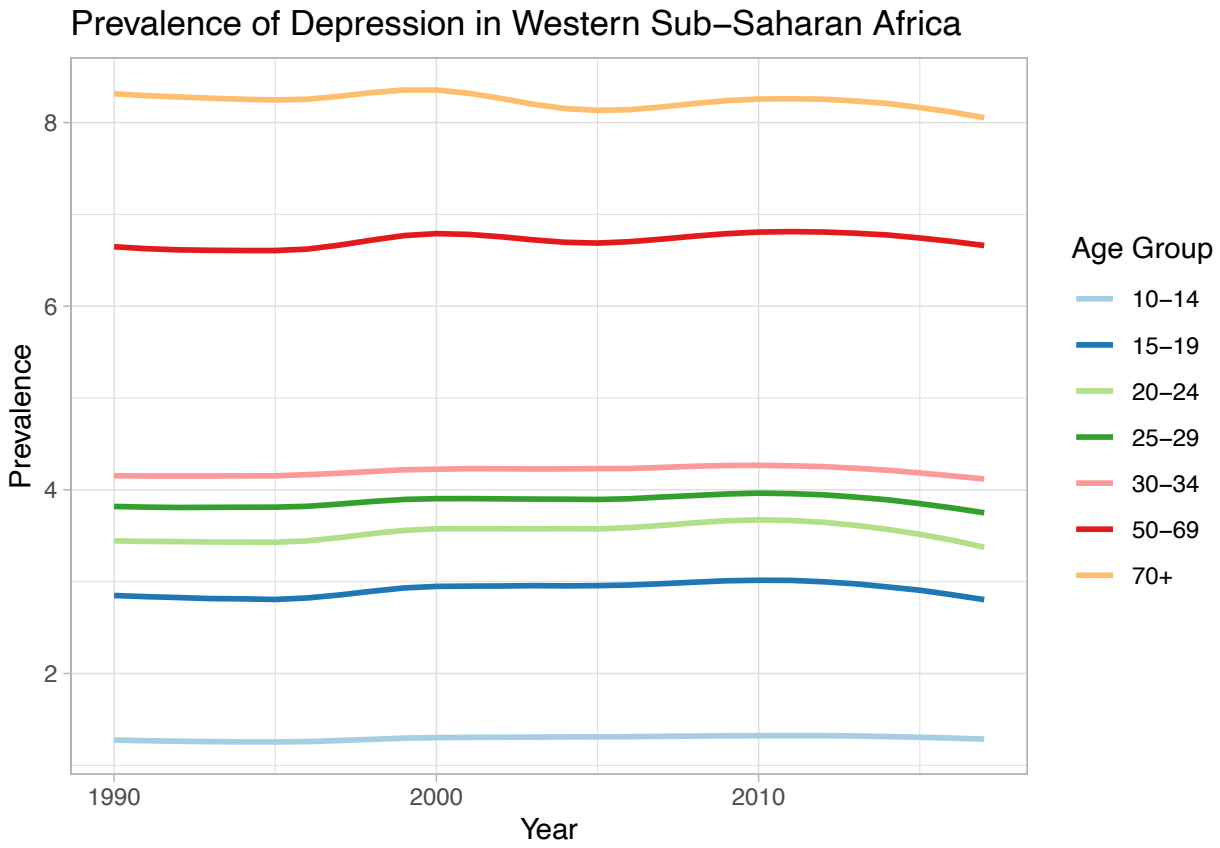

```
##pivot the data for Southern Sub-Saharan Africa
south_subSaharan_pivot <- pivot_my_data("Southern Sub-Saharan Africa")

##make the graph for pivoted data
graph_for_region("Southern Sub-Saharan Africa", south_subSaharan_pivot)
```




```
##pivot the data for Sub-Saharan Africa
subSaharan_pivot <- pivot_my_data("Sub-Saharan Africa")

##make the graph for pivoted data
graph_for_region("Sub-Saharan Africa", subSaharan_pivot)
```




```
##pivot the data for Western Sub-Saharan Africa
west_subSaharan_pivot <- pivot_my_data("Western Sub-Saharan Africa")

##make the graph for pivoted data
graph_for_region("Western Sub-Saharan Africa", west_subSaharan_pivot)
```



In each of the graphs above we notice that there is a big gap in the prevalence for the two older age groups and the middle-aged and younger age groups. The red and yellow lines correspond to the 50-60 and 70+ age groups, respectively, and they are set far apart from the rest of the age groups. Again, this trend is flipped from what we saw before with North America and Australasia.

There appears to be a difference between the groups, but is this difference significant? We can again create linear regression models that use age group to predict prevalence of depression. Here we're going to use just the Sub-Saharan African region because all of the graphs for the African regions look similar, and Sub-Saharan Africa seems like it might encompass all Sub-Saharan regions (west, east, etc.). This model should be representative of the trends occurring in these regions.

We can create a regression model to quantify the differences between prevalence of depression for the different age groups in the African regions. This will help us determine how the older age groups differ from some of the younger age groups. Let's first set the levels of the age group in the pivoted Sub Saharan dataframe so that we can make comparisons between the older age group and the 20-24 age group. The 20-24 group is within that younger and middle-age range, so this will give us a better comparison between the older and younger/middle age groups.

```
##relevel the age groups so 20-24 is the reference group
subSaharan_pivot$age_group <- factor(subSaharan_pivot$age_group,
                                                levels=c("20-24",
                                                         "15-19",
                                                         "10-14",
                                                         "25-29",
                                                         "30-34",
                                                         "50-69",
                                                         "70+" ))
##build linear model predicting prevalence with age group
```

```
subSaharan_mod <- lm(prevalence~age_group, data = subSaharan_pivot)
summary(subSaharan_mod)
```

```
##
## Call:
## lm(formula = prevalence ~ age_group, data = subSaharan_pivot)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.111548 -0.015114  0.003254  0.021853  0.089007
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.782016   0.006535  578.73   <2e-16 ***
## age_group15-19 -0.599884   0.009242  -64.91   <2e-16 ***
## age_group10-14 -2.379118   0.009242 -257.43   <2e-16 ***
## age_group25-29  0.215189   0.009242   23.28   <2e-16 ***
## age_group30-34  0.464627   0.009242   50.27   <2e-16 ***
## age_group50-69  2.965852   0.009242  320.91   <2e-16 ***
## age_group70+    4.351606   0.009242  470.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03458 on 189 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 1.187e+05 on 6 and 189 DF,  p-value: < 2.2e-16
```

In Sub-Saharan Africa there's a 3% increase in prevalence for the 50-69 age group compared to the 20-24 age group, and there's a 4.35% increase in prevalence for the 70+ age group compared to the 20-24 age group. The coefficients in the linear regression model are also all significant, so there are meaningful differences between prevalence of depression for older age groups compared to some of the younger age groups.

Overall, the general conclusion seems to be that we can't treat depression the same across regions if we're looking solely at age. We need to consider which region we're dealing with in order to determine which age group and type of treatment we should focus on. For example, if we're thinking about North America and Australasia, we should focus on the younger and middle-age groups and consider more reactive treatment. If depression is more prevalent among people aged 15-24 in these regions, then we need to react to it and try t fix it more immediately because it's happening pretty quickly. On the other hand, if we're considering Africa regions, we might think about employing preventative treatment because prevalence of depression is much higher among older age groups.

We should consider different factors that might be influencing the data as well especially since it is based off of self-reporting. For instance, different regions have different views on mental health, so those regions that aren't as progressive with the topic might have lower self-reporting compared to other regions. In North America, we have made huge strides in acceptance of mental health disorders, so it may be that we have higher self-reporting than other regions of the world.

Furthermore, it would be important to look at the prevalence within the 35-49 age group since it wasn't specifically included in the datset as its own group. I suspect that prevalence might be pretty similar to closer age groups as we saw previously, but this is an age group that is typically forgotten about when we talk about mental health disorders. We often focus on younger and older age groups because they have significant life changes and events that affect psychological well-being. However, the 35-49 age group is equally important because their mental health impacts their children, and the impact of the workplace is also pretty prominent in this age group.

Lastly, it would be interesting to explore other variables that could be affecting depression and could interact with age. Socioeconomic status is known to have a big impact on mental health, and we might find that it affects different age groups in different parts of the world differently. For example, different regions think about money differently, and some age groups are more concerned about money than others (adults more concerned than young children). Another variable we could consider is other mental health disorders within different age groups. Other mental heath disorders was included in a separate part of the bigger dataset, but being able to break that down by age group might illuminate further findings. For example, young females are known for struggling with eating disorders, which are usually linked to anxiety and depression. It would be interesting to discern whether there is an interaction between them and try to draw some conclusions about what other variables could impact prevalence among the different age groups.

Depression Prevalence Code

Sean McOsker

11/10/21

We are going to attempt to explore the geographic spread of prevalence of depression! Let's start with some libraries here

```
library(gganimate)
```

```
## Warning: package 'gganimate' was built under R version 4.0.5
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Let's read in our dataset

```
m <- read.csv("mentalhealthdata.csv")
```

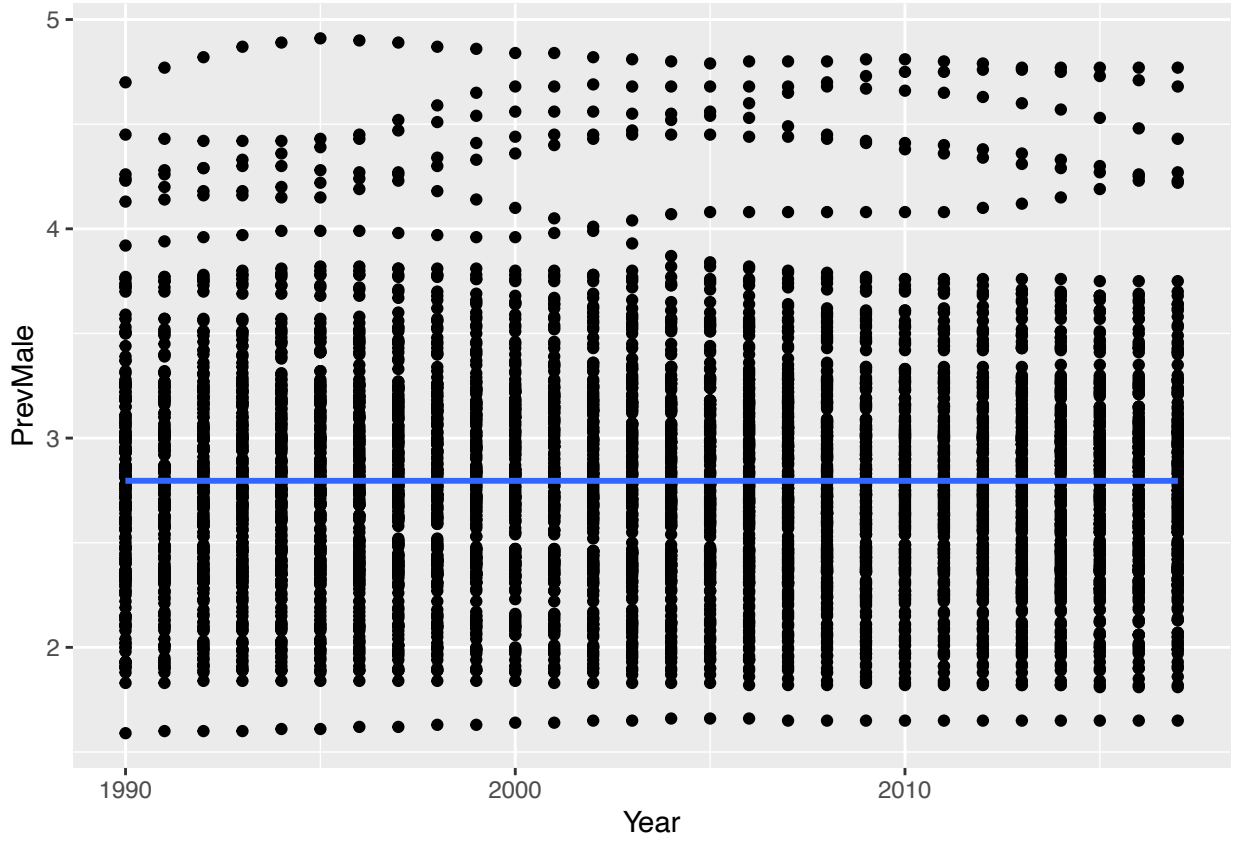
Some of our columns are not named intuitively. Let's change that.

```
m <- rename(m, PrevMale = X.VALUE.)
m <- rename(m, PrevFemale = X.VALUE..1)
```

Some basic graphical representations

```
pm<-ggplot(m, aes(Year, PrevMale)) + geom_point()
#with linreg
lm.m <- lm(PrevMale~Year, m)
pm + stat_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Not super helpful - it might be better for this interpretation to group all males from all countries together by the year...

Let's try that!

```
PrevMalebyYear <- m %>% group_by(Year) %>% summarise(mean(PrevMale), mean(PrevFemale))
PrevMalebyYear <- rename(PrevMalebyYear, Female = `mean(PrevFemale)`)
PrevMalebyYear <- rename(PrevMalebyYear, Male = `mean(PrevMale)`)
```

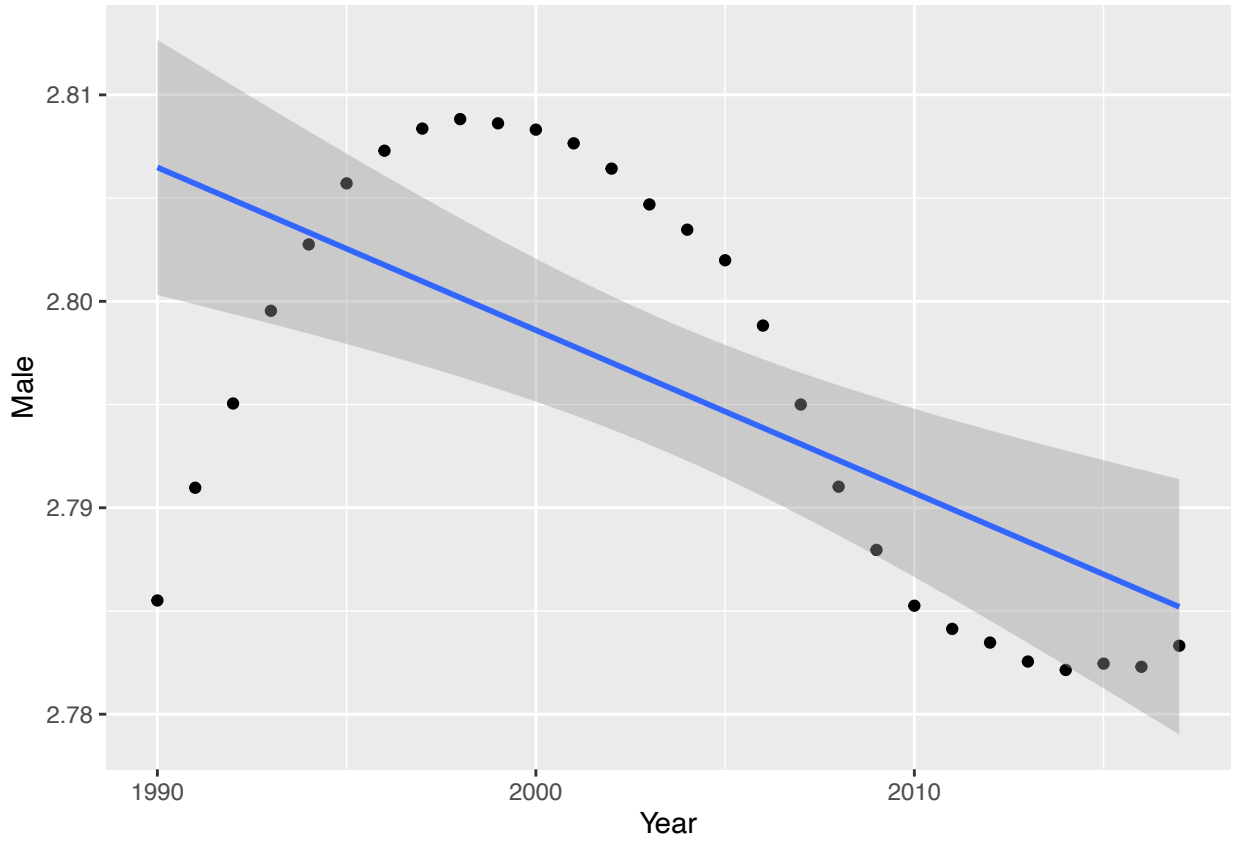
And perhaps likewise, let's find the average prevalence among males for all observations

```
PrevMaleOverall <- m%>% group_by(ï..Entity) %>% summarise(mean(PrevMale))
PrevMaleOverall <- rename(PrevMaleOverall, Country = ï..Entity)
PrevMaleOverall <- rename(PrevMaleOverall, Prev =`mean(PrevMale)`)
```

Let's now examine how these plots all run

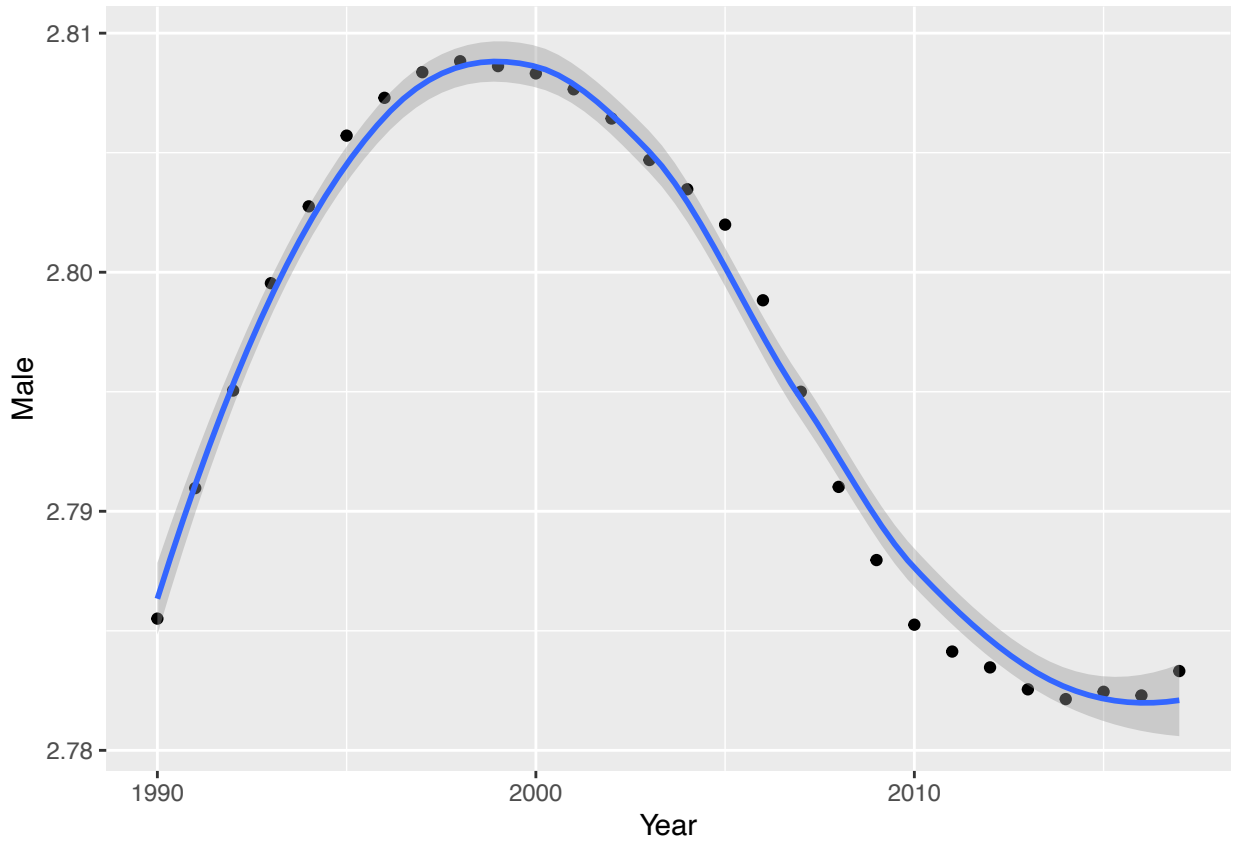
```
ggplot(PrevMalebyYear, aes(Year, Male)) + geom_point() + geom_smooth(method = "lm")
```

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




```
#and a loess method if we so choose
ggplot(PrevMalebyYear, aes(Year, Male)) + geom_point() + geom_smooth(method = "loess")
```

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



I'd like to get a good shapefile representation. Let's load in the "rgdal" package for using shapefiles

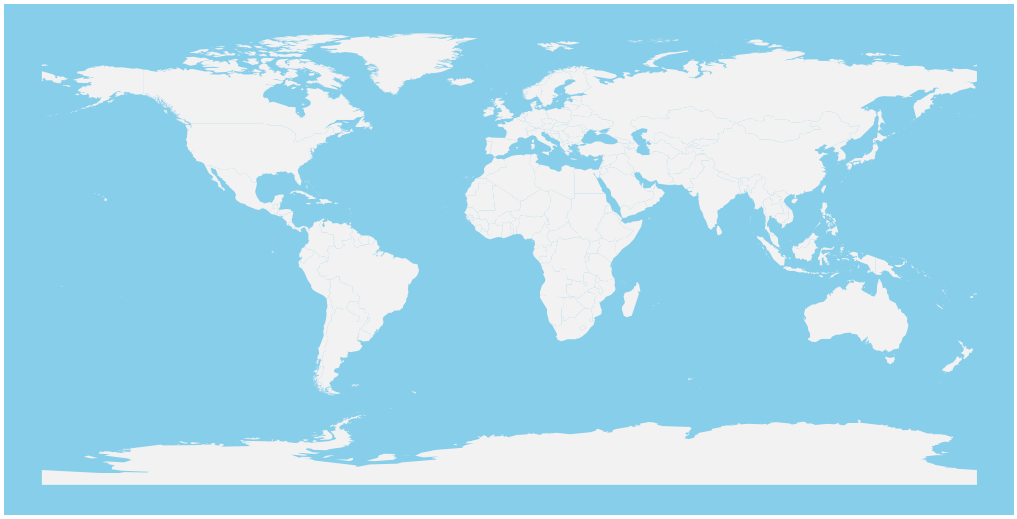
```
library(rgdal)
```

```
## Warning: package 'rgdal' was built under R version 4.0.5

## Loading required package: sp

## Warning: package 'sp' was built under R version 4.0.5

## Please note that rgdal will be retired by the end of 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
##
## rgdal: version: 1.5-27, (SVN revision 1148)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.2.1, released 2020/12/29
## Path to GDAL shared files: C:/Users/seanm/OneDrive/Documents/R/win-library/4.0/rgdal/gdal
## GDAL binary built with GEOS: TRUE
## Loaded PROJ runtime: Rel. 7.2.1, January 1st, 2021, [PJ_VERSION: 721]
## Path to PROJ shared files: C:/Users/seanm/OneDrive/Documents/R/win-library/4.0/rgdal/proj
## PROJ CDN enabled: FALSE
## Linking to sp version:1.4-5
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp or rgdal.
## Overwritten PROJ_LIB was C:/Users/seanm/OneDrive/Documents/R/win-library/4.0/rgdal/proj
```

```
wd <- getwd()
my_spdf <- readOGR(
  dsn= "C:/Users/seanm/OneDrive/Desktop/Grad School Stuff/TM_WORLD_BORDERS_SIMPL-0.3",
  layer="TM_WORLD_BORDERS_SIMPL-0.3",
  verbose=FALSE
)
```

Let's see what this shapefile looks like!

```
plot(my_spdf, col="#f2f2f2", bg="skyblue", lwd=0.25, border=0 )
```



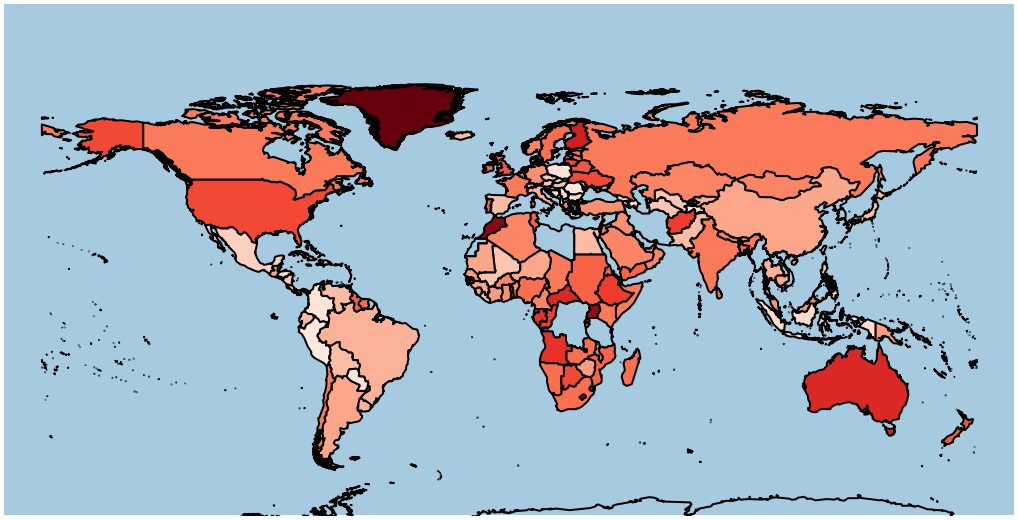
Let's merge our shapefile with our prevalence measures

```
world_merged <- merge(my_spdf, PrevMaleOverall, by.x = "NAME", by.y = "Country")
```

Alright, let's make it pretty and see how it looks!

```
# Palette of 30 colors
library(RColorBrewer)
my_colors <- brewer.pal(9, "Reds")
my_colors <- colorRampPalette(my_colors)(30)

# Attribute the appropriate color to each country
class_of_country <- cut(world_merged@data$Prev, 30)
my_colors <- my_colors[as.numeric(class_of_country)]
```

```
# Make the plot
plot(world_merged , ylim=c(0,40), col=my_colors ,  bg = "#A6CAE0")
```

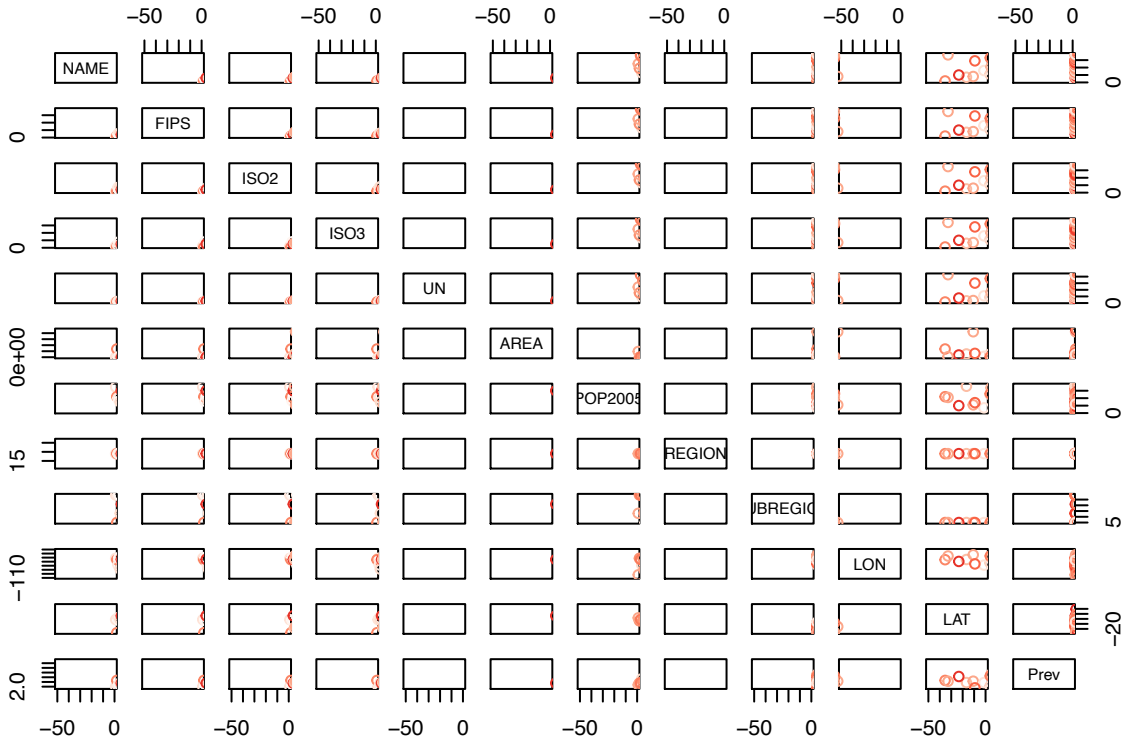


```
#legend(-119, 31.5, legend = levels(cut(world_merged@data$Prev, 30)), fill = my_colors, cex = 0.8, titl
```

Let's examine america specifically by referencing some regions

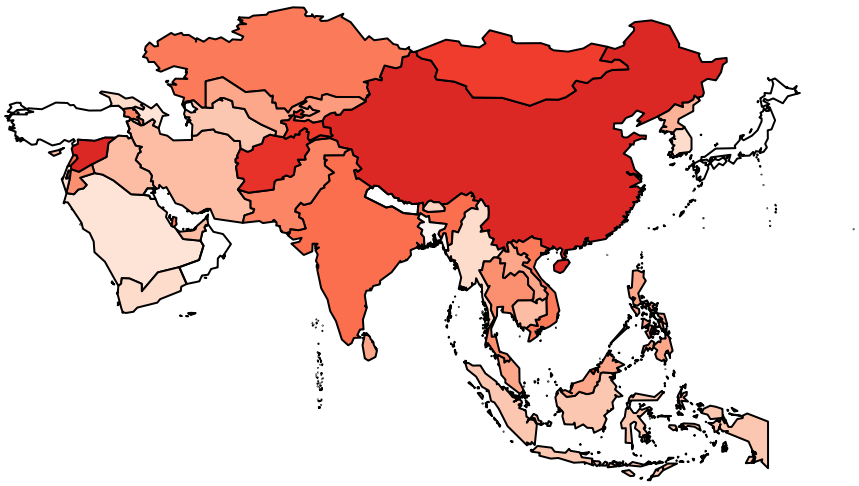
```
newdf <- merge(my_spdf@data, PrevMaleOverall, by.x = "NAME", by.y = "Country")
```

```
Americas <- newdf[newdf$REGION == 19, ]
plot(Americas, col = my_colors, xlim = c(-50, 0))
```

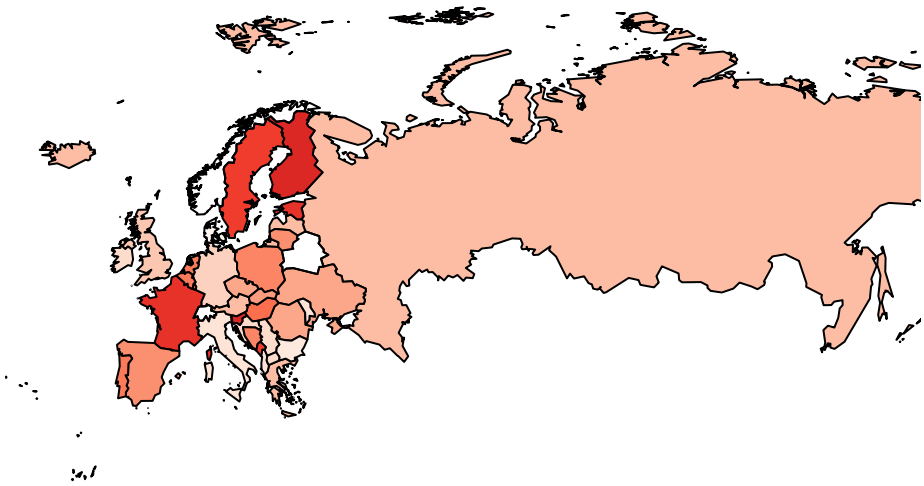


Now again for SE Asia and the Middle East

```
SEAsiaMiddleEast <- world_merged[world_merged@data$REGION == 142, ]
plot(SEAsiaMiddleEast, col = my_colors)
```



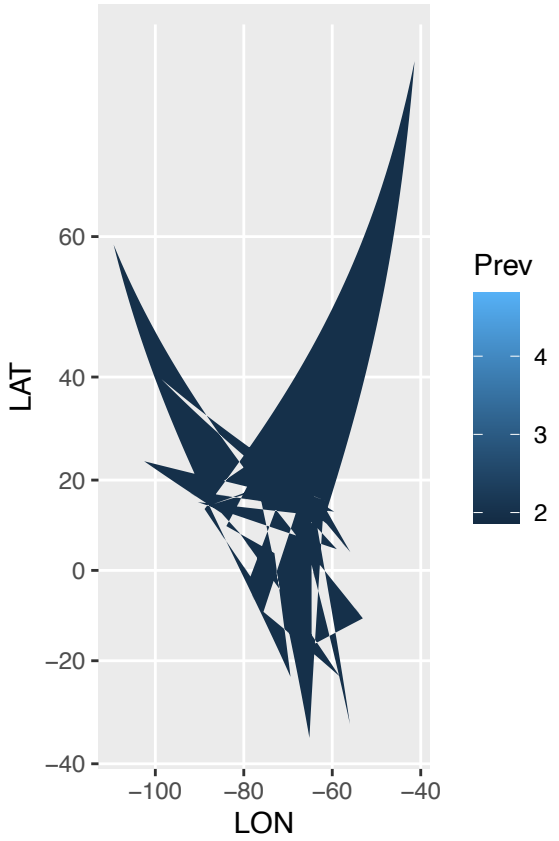
```
RussiaEurope <- world_merged[world_merged@data$REGION == 150, ]
plot(RussiaEurope, col = my_colors, xlim = c(0, 100))
```



Lets make another dataframe using the data from our SPDF

And let's visualize

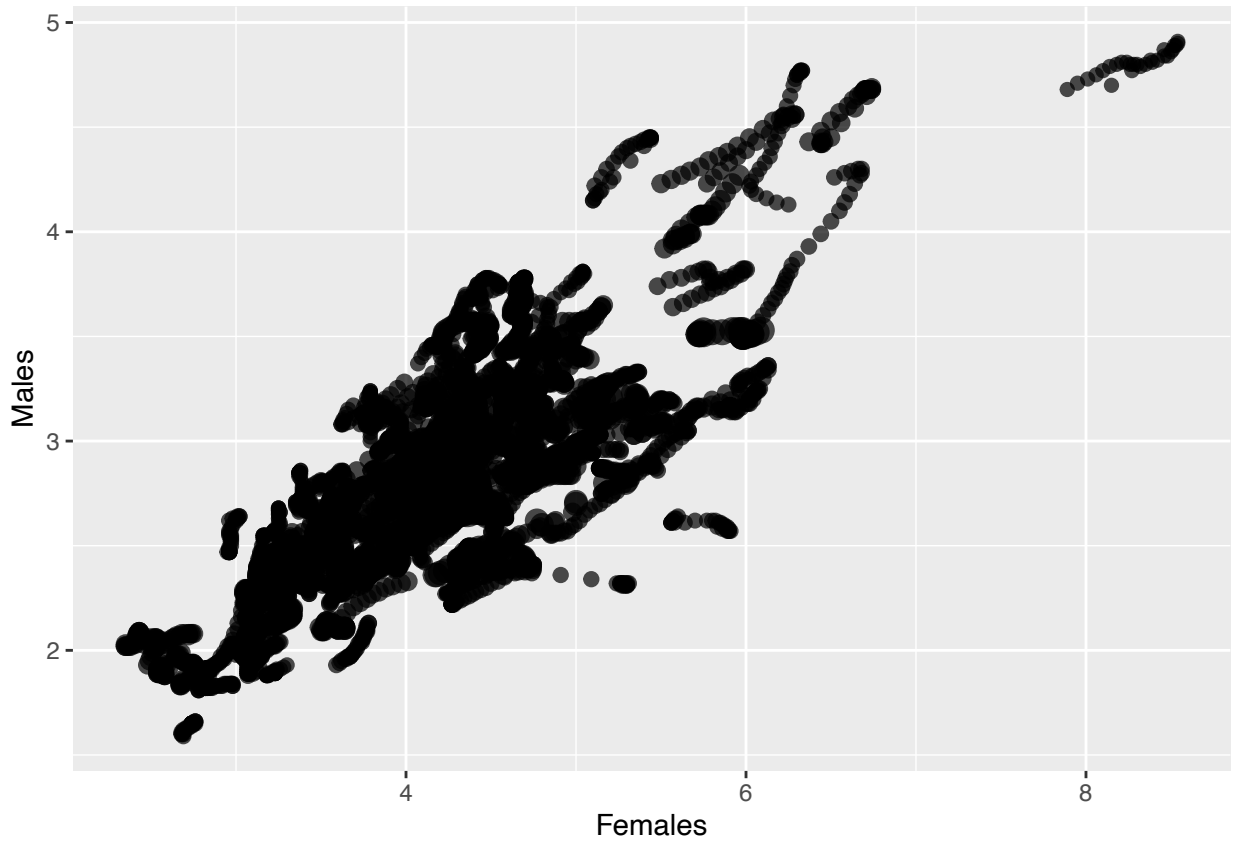
```
worldplot <- ggplot() +

  geom_polygon(data = Americas, aes(fill = Prev,
                                    x= LON,
                                    y = LAT,
                                    )) + coord_map()

worldplot
```



Now let's get fun! Time to make some animations.

```
p <- ggplot(m,
            aes(x = m$PrevFemale, y = m$PrevMale, size = m$Population)) +
  geom_point(show.legend = FALSE, alpha = 0.7) +
  scale_color_viridis_d() +
  scale_size(range = c(2, 12)) +
  labs(x = "Females", y = "Males")

p
```

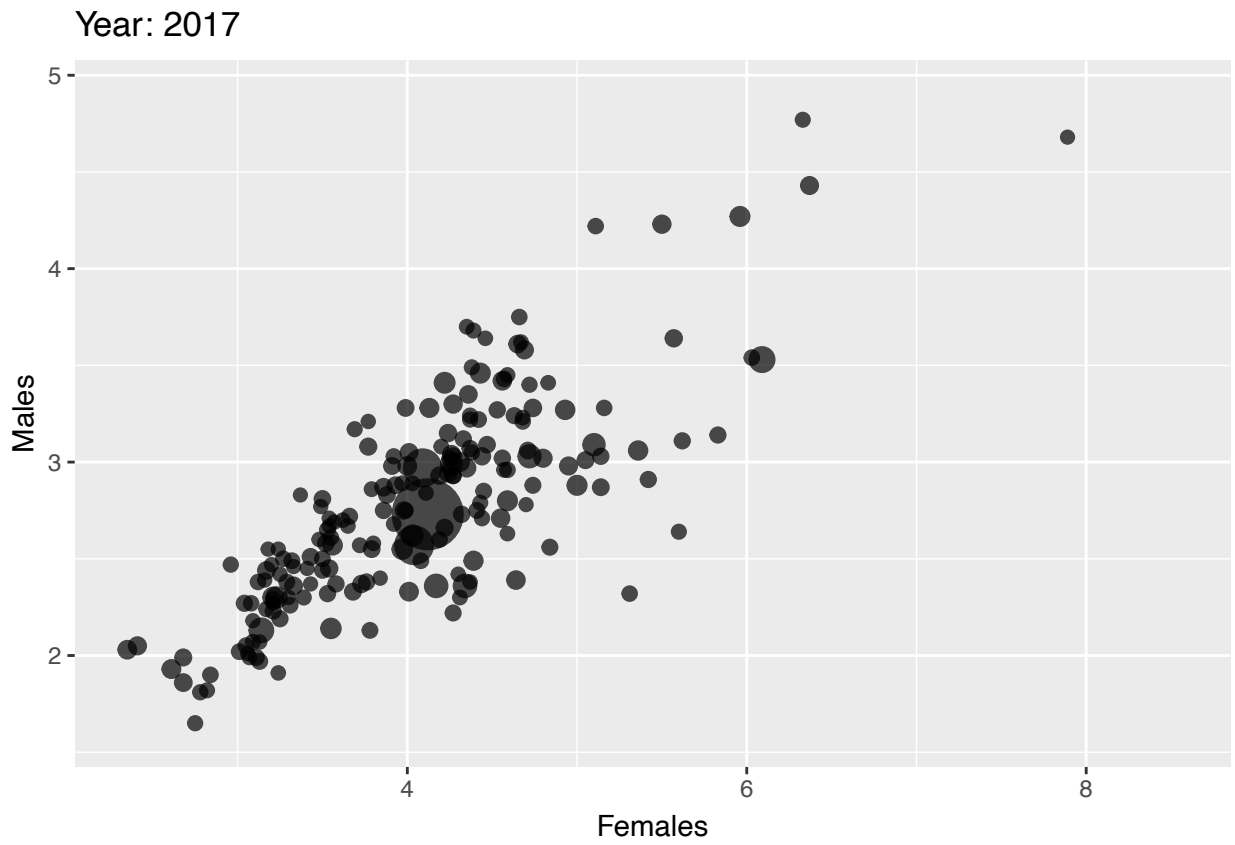
```
## Warning: Use of ‘m$PrevFemale‘ is discouraged. Use ‘PrevFemale‘ instead.

## Warning: Use of ‘m$PrevMale‘ is discouraged. Use ‘PrevMale‘ instead.

## Warning: Use of ‘m$Population‘ is discouraged. Use ‘Population‘ instead.
```



Adding animation over the year variable

```
p +transition_time(m$Year) + labs(title = "Year: {frame_time}")
```

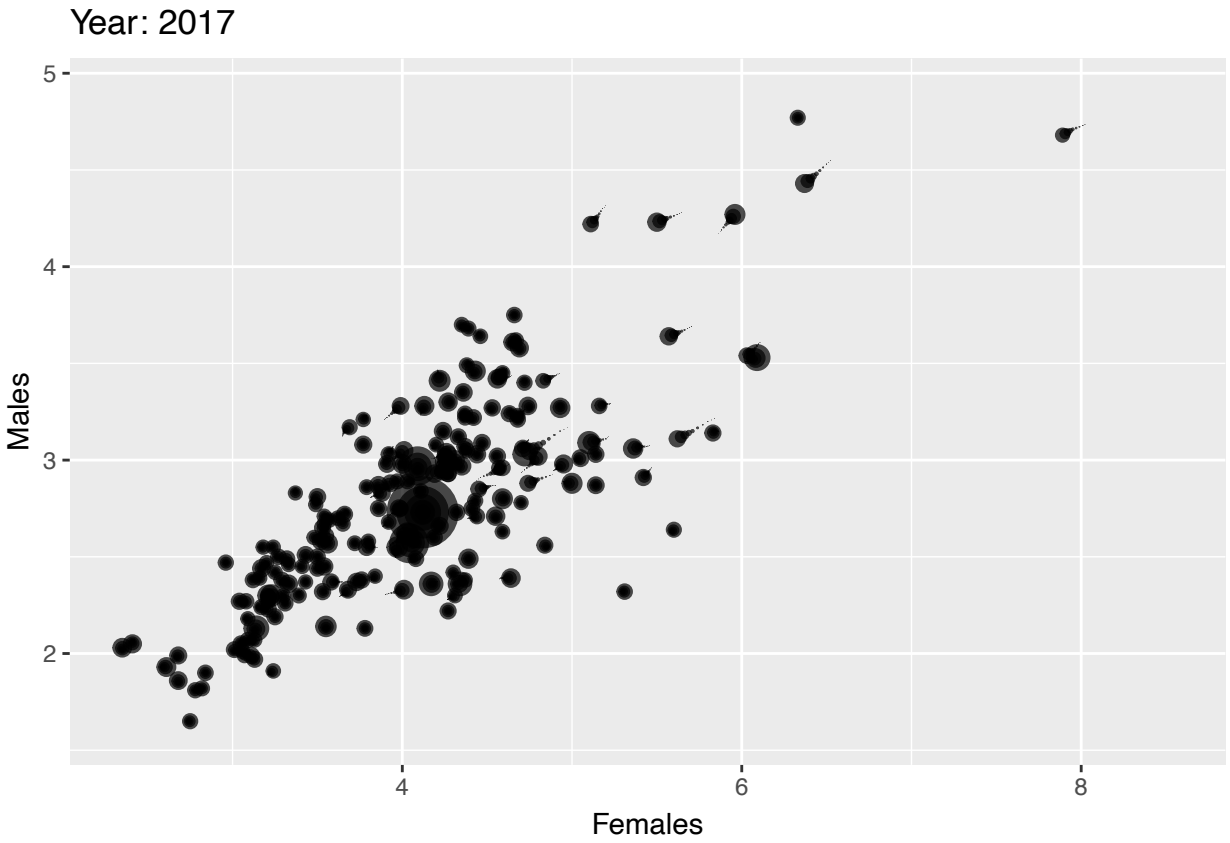
```
## Warning: Use of ‘m$PrevFemale‘ is discouraged. Use ‘PrevFemale‘ instead.

## Warning: Use of ‘m$PrevMale‘ is discouraged. Use ‘PrevMale‘ instead.

## Warning: Use of ‘m$Population‘ is discouraged. Use ‘Population‘ instead.
```



Now let's add some tails to better see changes

```
pwithtails<-p + transition_time(m$Year) + labs(title = "Year: {frame_time}") +
  shadow_wake(wake_length = 0.1, alpha = FALSE)

pwithtails
```

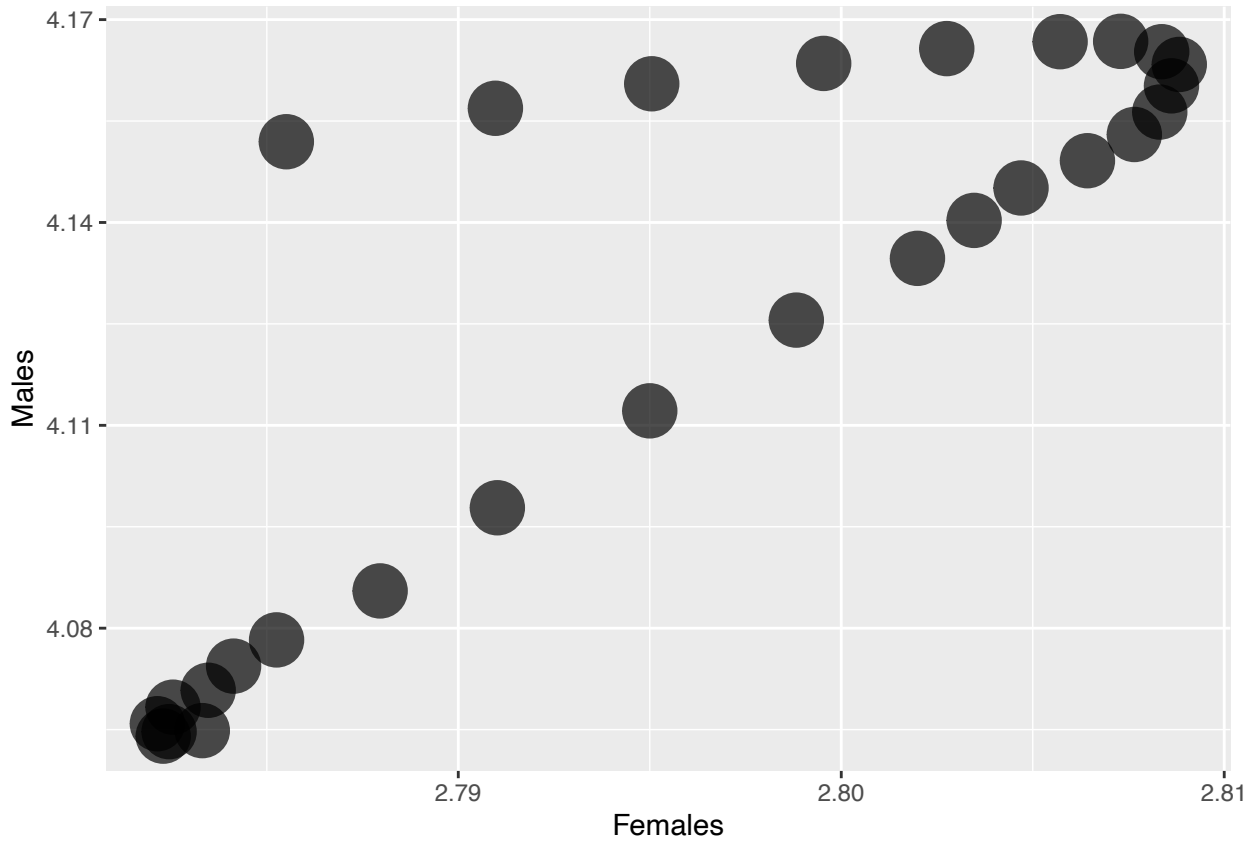
```
## Warning: Use of ‘m$PrevFemale‘ is discouraged. Use ‘PrevFemale‘ instead.

## Warning: Use of ‘m$PrevMale‘ is discouraged. Use ‘PrevMale‘ instead.

## Warning: Use of ‘m$Population‘ is discouraged. Use ‘Population‘ instead.
```




```
#anim_save(file = "bycountry", animation = pwithtails, path = "C:/Users/seanm/OneDrive/Desktop/Grad Sch
```

Maybe map this as females vs. Males?

```
poverall <- ggplot(PrevMalebyYear,
            aes(x = Male , y = Female, size = 2)) +
  geom_point(show.legend = FALSE, alpha = 0.7) +
  scale_color_viridis_d() +
  scale_size(range = c(2, 12)) +
  labs(x = "Females", y = "Males")

poverall
```



And make it an anim and save it!

```
#overall<-poverall + transition_time(PrevMalebyYear$Year) + labs(title = "Year: {frame_time}") +
 # shadow_wake(wake_length = 0.2, alpha = FALSE)

#anim_save(file = "overall", animation = overall, path = "C:/Users/seanm/OneDrive/Desktop/Grad School S
```

Let's try one more dataset which better illustrates region differences

```
finalmerged <- merge(PrevMalebyYear, my_spdf)
```

And renaming it!

```
m <- rename(m, Country = ï..Entity)
```

```
#fullspdf <- merge(my_spdf, m, by.x = "NAME", by.y = "Country", all.x = TRUE)
```

Education level and employment status data analysis

Emma Wagner

11/15/2021

In this R markdown, I will analyze the education and employment dataset. To start, I loaded the necessary libraries and read in the excel sheet. I used the head() function as well to make sure the data was read in correctly.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tibble' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

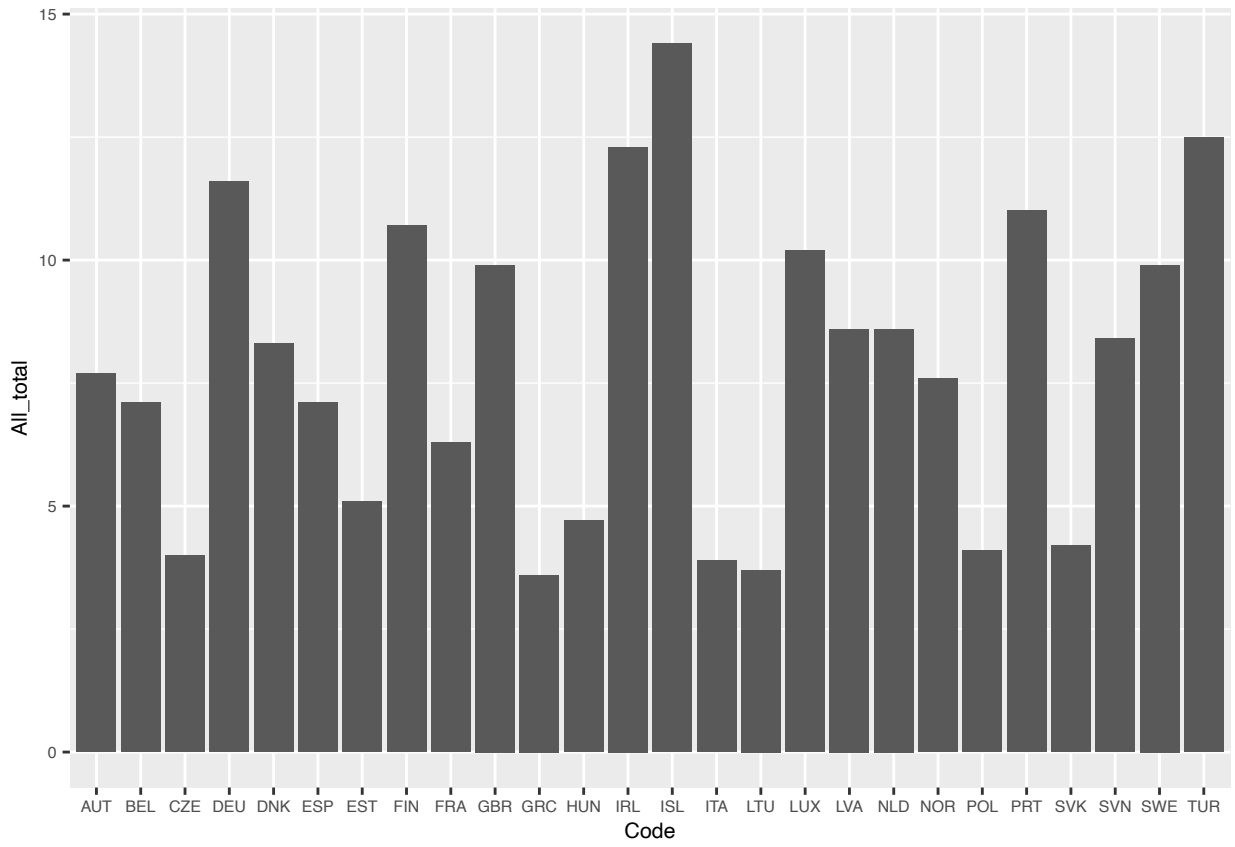
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)

edu = read_excel("/Users/emmawagner/Desktop/Data Wrangling/Final project/Mental health data cleaned.xlsx
head(edu)
```

```
## # A tibble: 6 x 15
##   Entity         Code   Year All_active All_employed All_total BelowUS_active
##   <chr>          <chr> <dbl>      <dbl>        <dbl>     <dbl>           <dbl>
## 1 Austria        AUT    2014        6.5          4.7       7.7            15.5
## 2 Belgium        BEL    2014        5            4.1       7.1             7.1
## 3 Czech Republic CZE    2014        3            2.6       4               2.1
## 4 Denmark        DNK    2014        6.7          5.7       8.3            10.4
## 5 Estonia        EST    2014        3.8          3.8       5.1             4.7
## 6 Finland        FIN    2014        8.5          7.2      10.7             7.4
## # ... with 8 more variables: BelowUS_employed <dbl>, BelowUS_total <dbl>,
## #   Tertiary_active <dbl>, Tertiary_employed <dbl>, Tertiary_total <dbl>,
## #   US_active <dbl>, US_employed <dbl>, US_total <dbl>
```

Next, I created a preliminary bar plot to look at the data. I created a bar graph with country on the x axis and the depression rate (%) on the y axis.

```
bar = ggplot(edu, aes(x= Code, y = All_total)) +
  geom_bar(stat= "identity") +
  theme(text = element_text(size=8))
bar
```



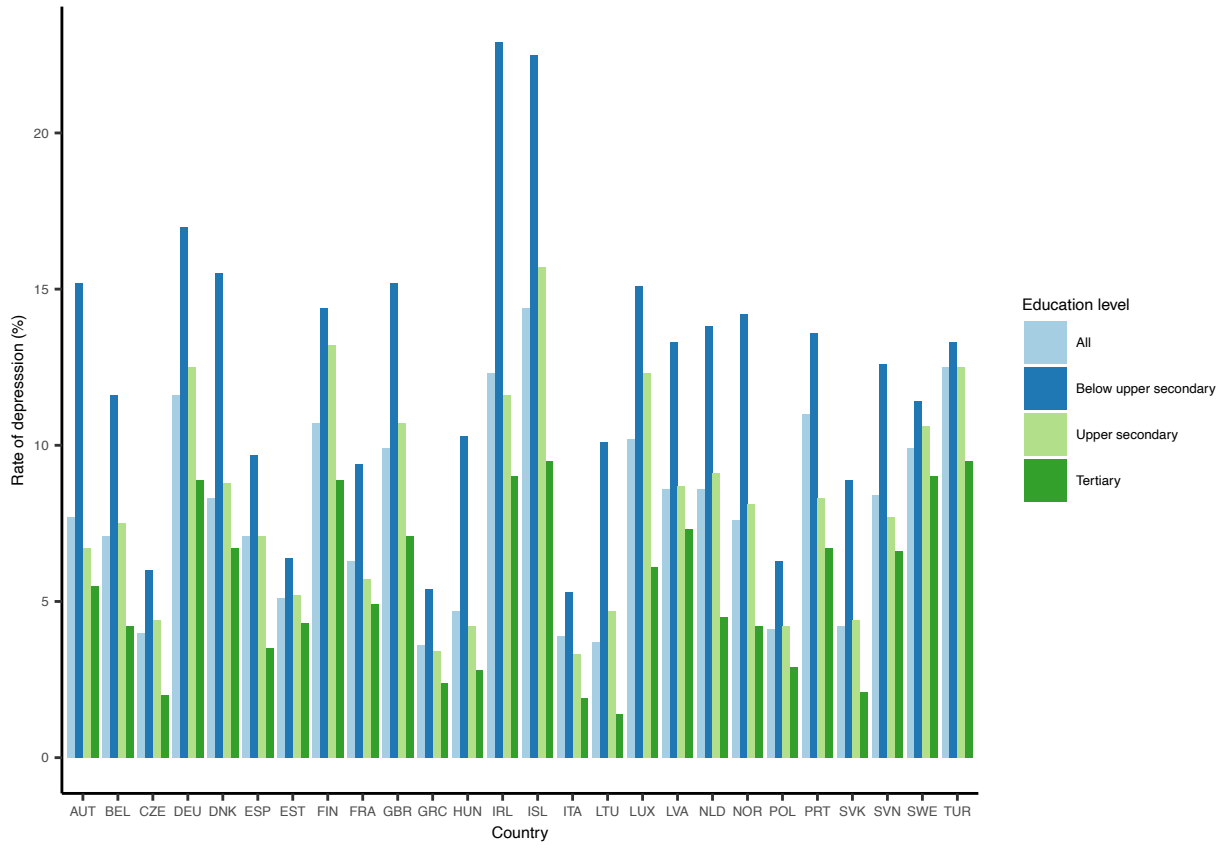
Here we can see there is a range in depression rates between different countries- however this is not very useful for analyses of education and employment. In order to better visualize and analyze the data using the tidyverse, we need to convert the data into long form.

```
# Put data into long format
edu_long = pivot_longer(edu, cols = c("All_active", "All_employed", "All_total", "BelowUS_active", "Belo
                        "Tertiary_active", "Tertiary_employed", "Tertiary_total", "US_active", "US_emplo
                        names_to = "EduLevel")

# Split EduLevel into two columns- EduLevel and Employment
edu_long = separate(edu_long, col = EduLevel, into=c("EduLevel", "Employment"), sep = "_")
```

edu_long now contains the data in long form, with three new columns. EduLevel and Employment contains the education level and employment connected to each depression rate. The value column contains the rate of depression. This reduces the number of variables to 6, from 15 previously.

Next, I created a barplot of data from all of the countries to do some initial analysis of education. I filtered the data to include only the total data for employment.

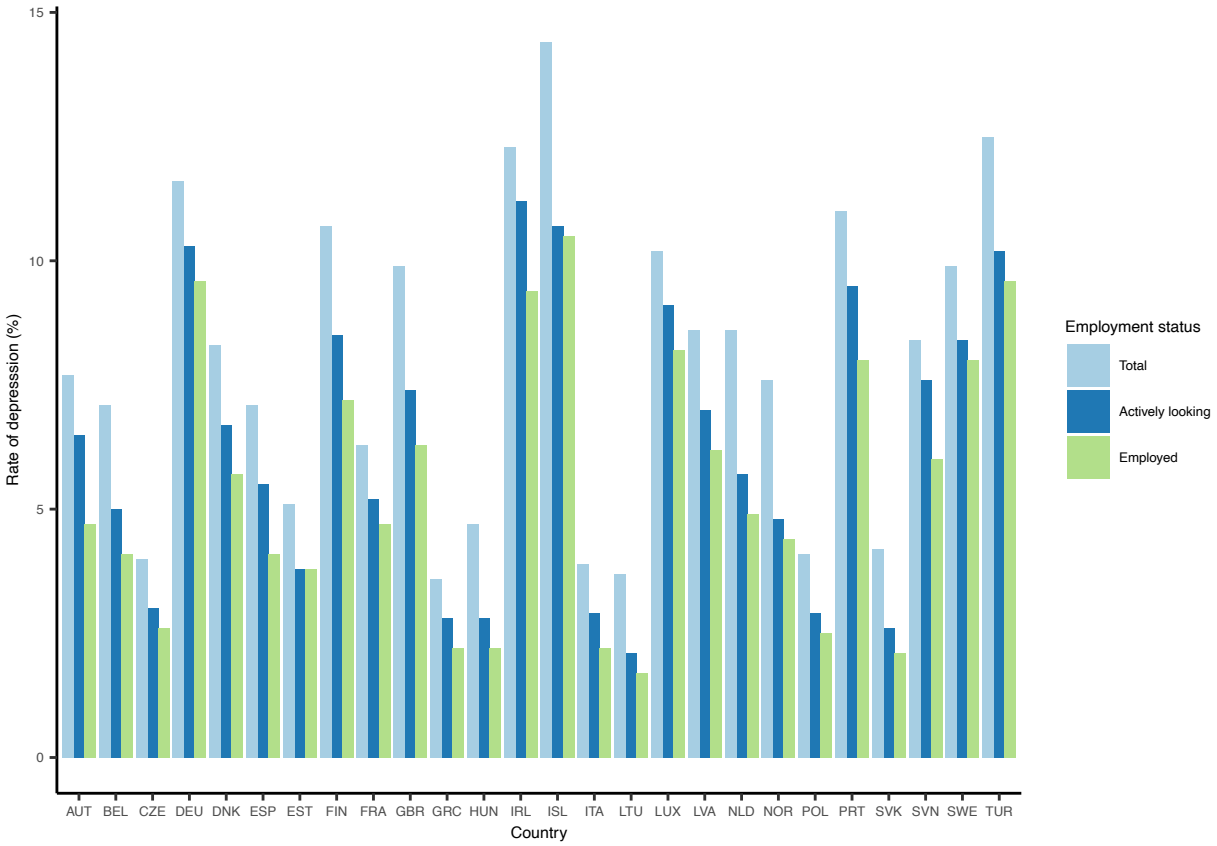
```
# barplot by countries
level_order = c("All", "BelowUS", "US", "Tertiary")
totalonly= edu_long %>%
  filter(Employment == "total")
bar = ggplot(totalonly, aes(x= Code, y = value, fill = factor(EduLevel, level= level_order))) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_classic() +
  scale_fill_brewer(palette = "Paired", labels = c("All", "Below upper secondary", "Upper secondary", "T
  theme(text = element_text(size=6)) +
  labs(y= "Rate of depresssion (%)", x = "Country", fill = "Education level")
bar
```



In every country, the highest rates of depression were among people with a below upper secondary level of education and the lowest rates of depression were for people with a tertiary level education. This indicates that education reduces depression rates.

Next, I did the same thing for employment data, filtering to include all levels of education.

```
# barplot by countries
level_order = c("total", "active", "employed")
allonly= edu_long %>%
  filter(EduLevel == "All")
bar = ggplot(allonly, aes(x= Code, y = value, fill = factor(Employment, level= level_order))) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_classic() +
  scale_fill_brewer(palette = "Paired", labels = c("Total", "Actively looking", "Employed")) +
  theme(text = element_text(size=6)) +
  labs(y= "Rate of depresssion (%)", x = "Country", fill = "Employment status")
bar
```



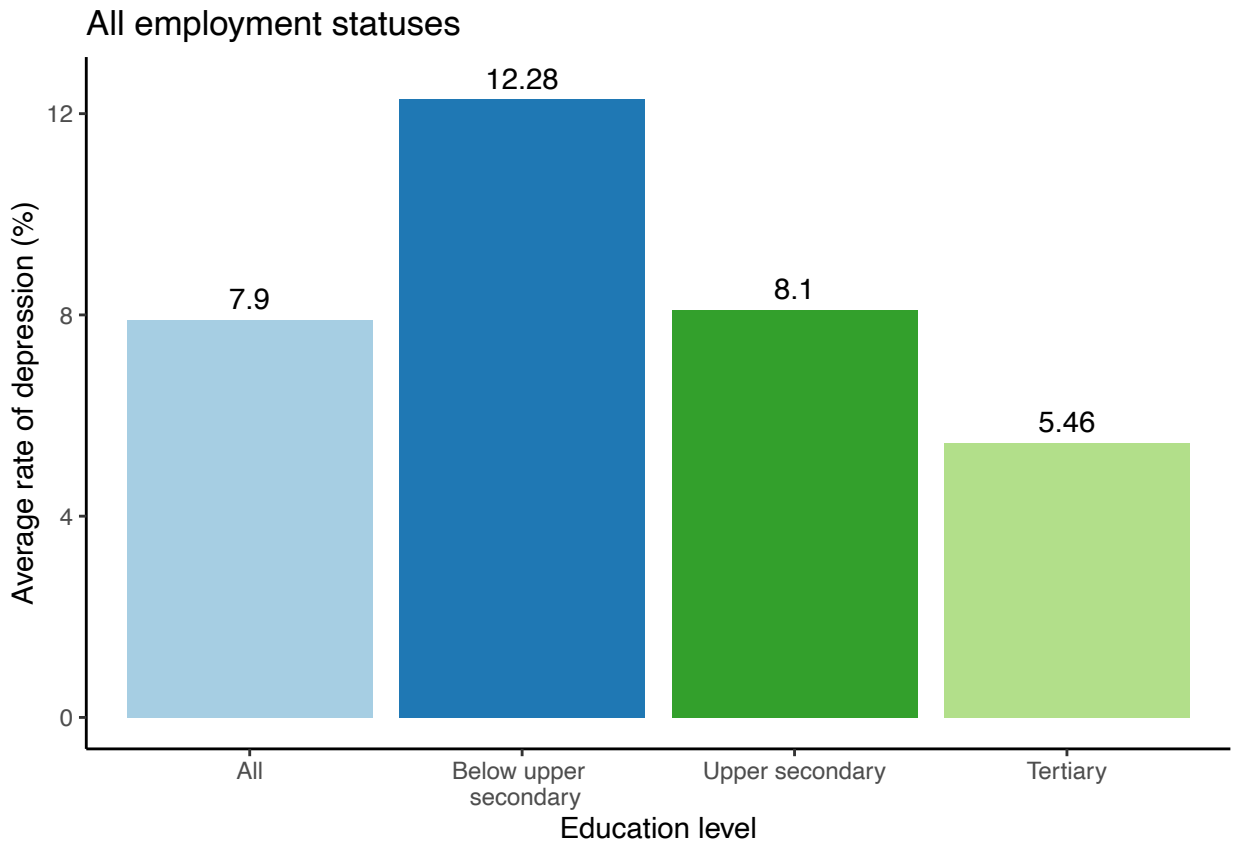
Every country follows the same pattern- employed people have the lowest rates of depression, people actively looking for work have a slightly higher rate, and the total population (including unemployed people) had the highest rates. This supports the idea that employment reduces rates of depression.

In these two graphs, there is a significant level of variation in depression rates between countries. My group mates analyzed depression prevalence at the regional and country level with bigger data sets, so I am not going to focus on differences between countries, just on differences between groups based on education and employment.

With 26 countries the above visualizations are a bit cluttered, so in the next plots I will aggregate the data.

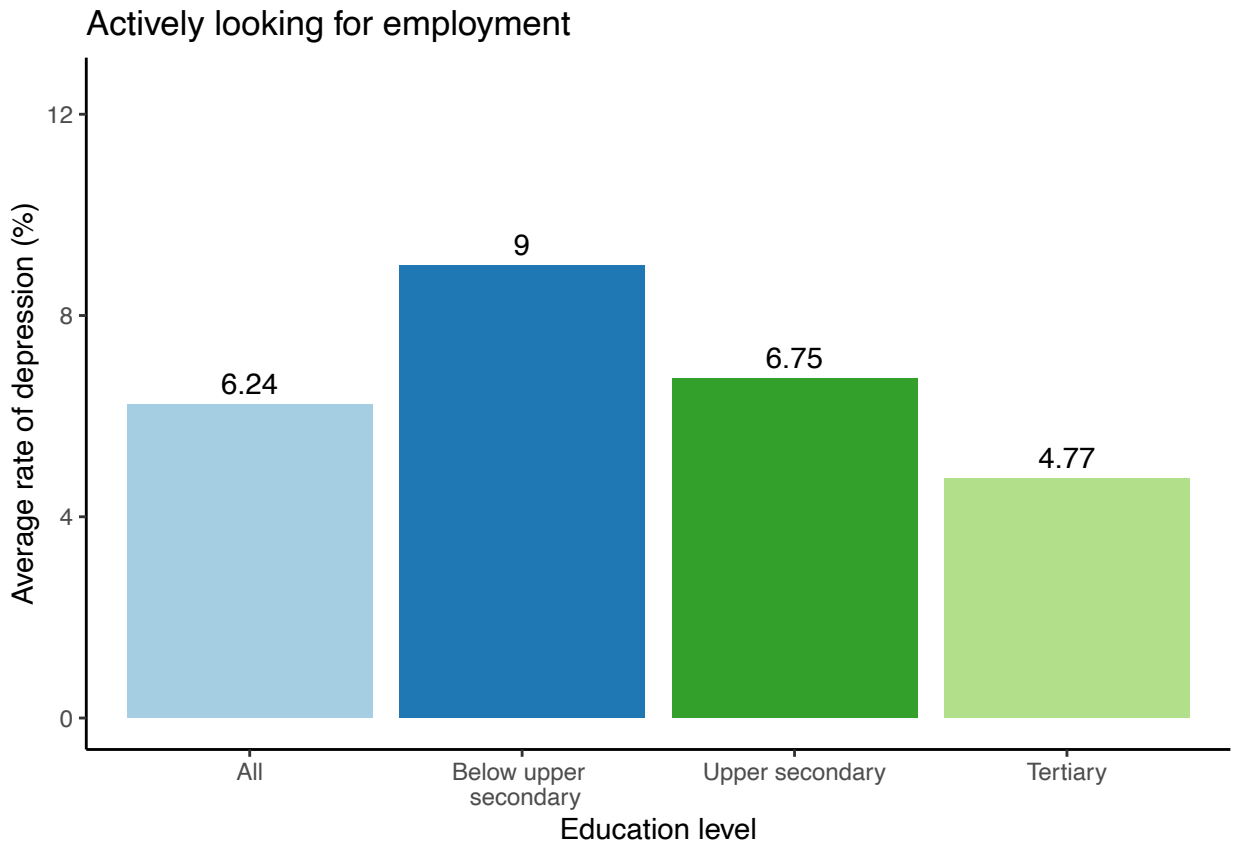
Below, I filtered the data by employment level to create plots of the average depression rate for that group. Each plot shows the depression rates for different levels of education for that employment status.

```
level_order = c("All", "BelowUS", "US", "Tertiary")
# for total data
bar2 =  edu_long %>%
  filter(Employment == "total") %>%
  group_by(EduLevel) %>%
  summarise(meanByEdu = mean(value)) %>%
  ggplot(aes(x= factor(EduLevel, level= level_order), y = meanByEdu, fill = EduLevel)) +
  geom_col() +
  scale_y_continuous(limits =c(0, 12.5)) +
  theme_classic() +
  geom_text(aes(label = (round(meanByEdu, digits = 2)), vjust = -0.5)) +
  labs(x= "Education level", y= "Average rate of depression (%)", title = "All employment statuses") +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Paired") +
  scale_x_discrete(labels = c("All", "Below upper \n secondary", "Upper secondary", "Tertiary"))
bar2
```

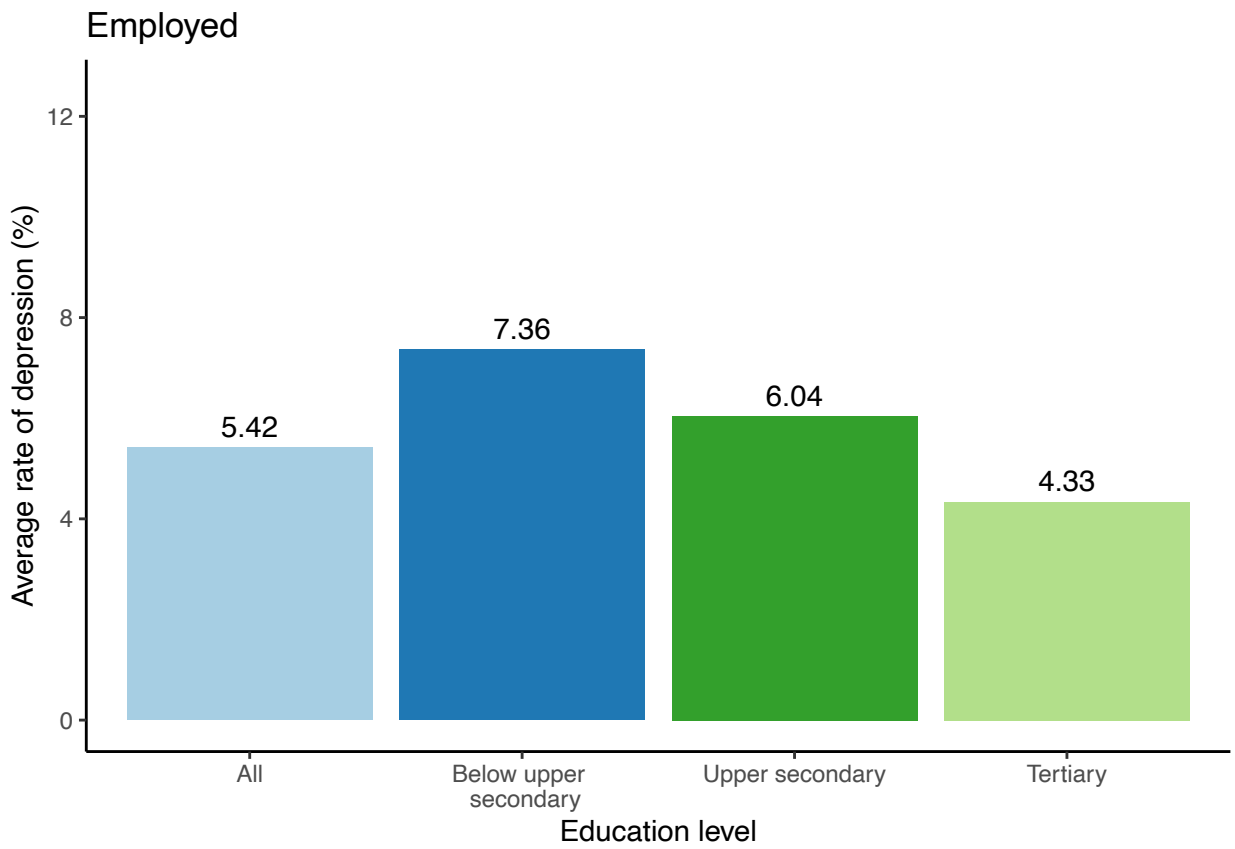



```
# for active data
bar3 =  edu_long %>%
  filter(Employment == "active") %>%
  group_by(EduLevel) %>%
  summarise(meanByEdu = mean(value)) %>%
  ggplot(aes(x= factor(EduLevel, level= level_order), y = meanByEdu, fill = EduLevel)) +
```

```
  geom_col() +
  scale_y_continuous(limits =c(0, 12.5)) +
  geom_text(aes(label = (round(meanByEdu, digits = 2)), vjust = -0.5)) +
  theme_classic() +
  labs(x= "Education level", y= "Average rate of depression (%)", title = "Actively looking for employm
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Paired") +
  scale_x_discrete(labels = c("All", "Below upper \n secondary", "Upper secondary", "Tertiary"))
bar3
```




```
# for employed
bar3 =  edu_long %>%
  filter(Employment == "employed") %>%
  group_by(EduLevel) %>%
  summarise(meanByEdu = mean(value)) %>%
  ggplot(aes(x= factor(EduLevel, level= level_order), y = meanByEdu, fill = EduLevel)) +
  geom_col() +
  scale_y_continuous(limits =c(0, 12.5)) +
  geom_text(aes(label = (round(meanByEdu, digits = 2)), vjust = -0.5)) +
  theme_classic() +
  labs(x= "Education level", y= "Average rate of depression (%)", title = "Employed") +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Paired") +
  scale_x_discrete(labels = c("All", "Below upper \n secondary", "Upper secondary", "Tertiary"))
bar3
```

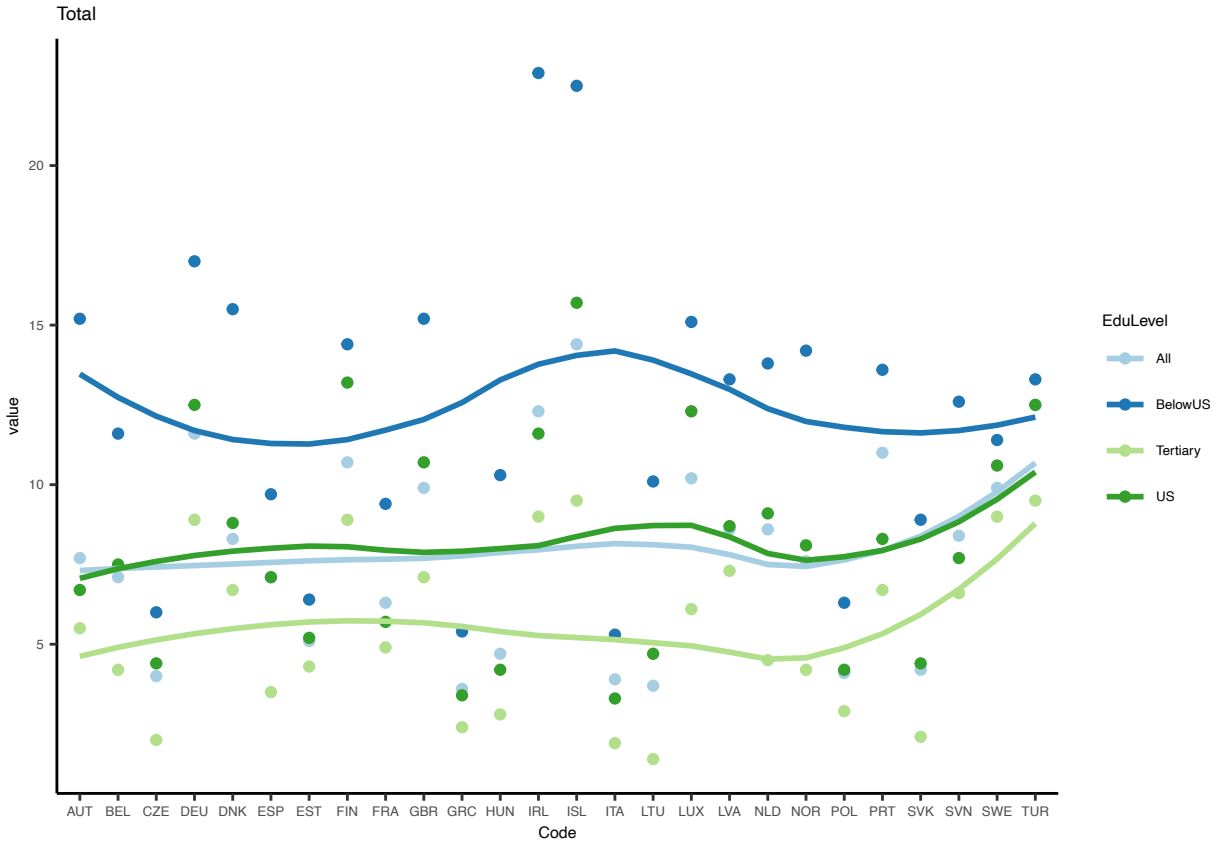


These plots show the same patterns as the country level plots above: across all levels of employment, the more educated groups had lower average rates of depression than less educated groups. Similarly, employed people had lower average rates of depression than both those actively looking for work and the total population. Overall, the lowest average rate of depression (4.33%) was for the employed and tertiary education groups. The highest average rate (12.28%) was for below upper secondary education in the total population.

For a final set of plots, I plotted scatterplots with smoothing lines to show the trends across countries for another visual display. Each plot is for a different employment status

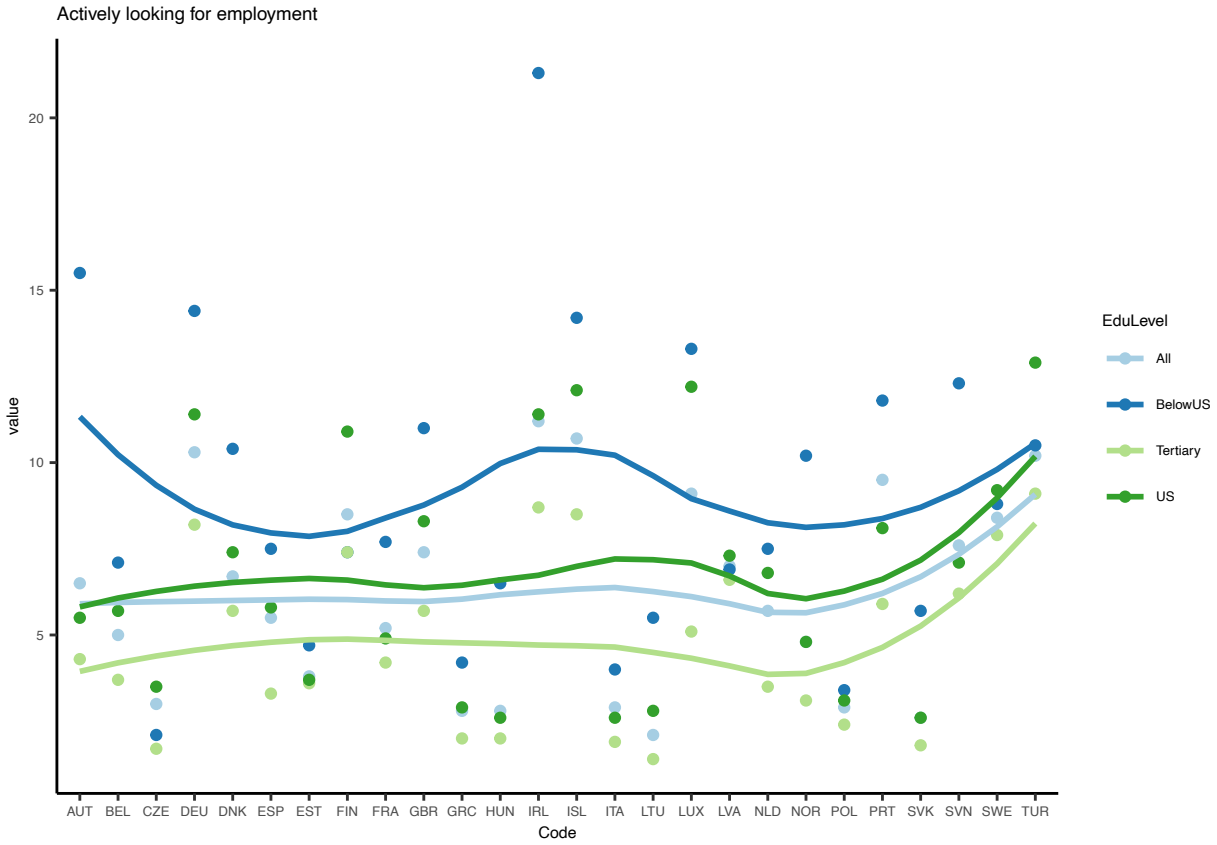
```
# Scatter plot with smoothing lines for total employment data
all =edu_long %>%
  filter(Employment == "total")
p = ggplot(all, aes(x = Code, y = value, fill = EduLevel, color= EduLevel, group = EduLevel)) +
  geom_point()+
  geom_smooth(method = loess, se =FALSE) +
  scale_color_brewer(palette = "Paired") +
  theme_classic() +
  theme(text = element_text(size=6)) +
  labs(title = "Total")
p
```

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



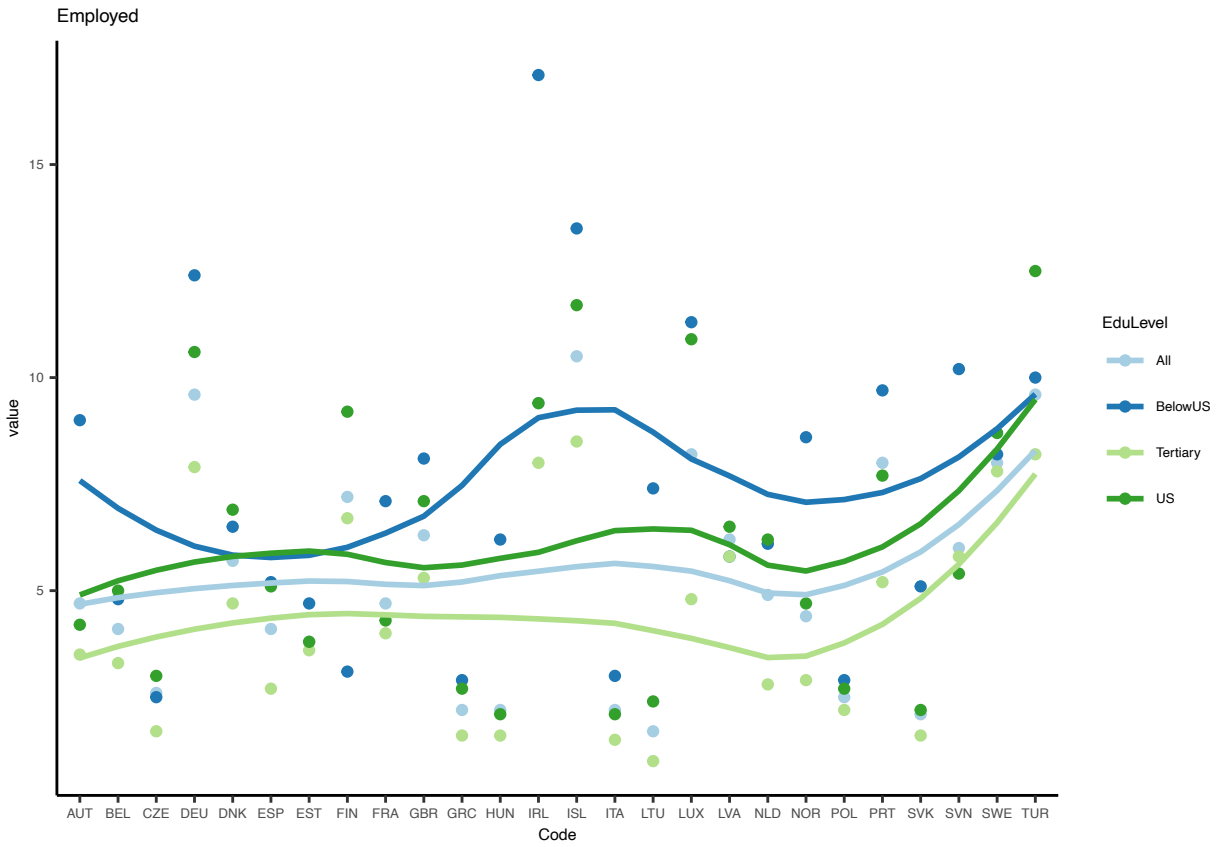

```
# Scatter plot with smoothing lines for active
active =edu_long %>%
  filter(Employment == "active")
p = ggplot(active, aes(x = Code, y = value, fill = EduLevel, color= EduLevel, group = EduLevel)) +
  geom_point()+
  geom_smooth(method = loess, se =FALSE) +
  scale_color_brewer(palette = "Paired") +
  theme_classic() +
  theme(text = element_text(size=6)) +
  labs(title = "Actively looking for employment")
p
```

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




```
# Scatter plot with smoothing lines for employed
employ =edu_long %>%
  filter(Employment == "employed")
p = ggplot(employ, aes(x = Code, y = value, fill = EduLevel, color= EduLevel, group = EduLevel)) +
  geom_point()+
  geom_smooth(method = loess, se =FALSE) +
  scale_color_brewer(palette = "Paired") +
  theme_classic() +
  theme(text = element_text(size=6)) +
  labs(title = "Employed")
p
```

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



Overall, this showed the same trends as the plots above. More educated people had lower levels of depression and less educated people had higher rates. Because of the amount of variation between countries, this plot is not quite as effective at visualizing the data as the bar plots above.

Next, I ran some statistical tests on the data to see if the differences between groups are statistically significant.

First, I ran a one-way anova to see if there were significant differences between groups based on education levels.

```
summary(aov(value~EduLevel, data = edu_long))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## EduLevel      3    884  294.55   23.89 6.27e-14 ***
## Residuals   308   3797   12.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value$<$ 0.05, so the test is statistically significant and we can reject the null hypothesis of no differences between groups.

Next, I ran a one way anova based on employment

```
summary(aov(value~Employment, data = edu_long))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Employment    2    377  188.44   13.53 2.33e-06 ***
```

```
## Residuals   309   4304   13.93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, the p value is less than .05 we can reject the null hypothesis, indicating that the differences between groups are statistically significant.

Finally, I ran a two way anova with both EduLevel and Employment.

```
summary(aov(value~Employment +EduLevel, data = edu_long))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Employment    2    377  188.44   16.86 1.13e-07 ***
## EduLevel      3    884  294.55   26.35 3.44e-15 ***
## Residuals   306   3420   11.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both p values are less than 0.05 indicating the differences are statistically significant.

Overall, the data showed clearly the positive effects of education and employment on reducing depression rates. This highlights how issues such as the increasing unaffordability of college, unemployment, and lack of quality public school educations for all students are detrimental to mental health at the population level.

In future analyses, it would be interesting to analyze a dataset with more countries to see if these trends hold true in other countries and regions. Initially, I was not expecting education to reduce depression rates so significantly. In the United States, there is a pretty unhealthy culture around work with long hours with little vacation time. So, I was expecting that people with higher levels of education would be more likely to have high stress jobs and would have rates of depression similar to those with less education. On the other hand, in Europe, where this data is from, there seems to be a more balanced culture around work, with more time off, paid parental leave, etc. I'm curious if this impacts depression rates. So, it would be interesting to analyze how culture and policies around employment and education influence depression rates. Additionally, this data was only from 2014. Analyzing data from other years could allow for more complex analyses and allow us to analyze trends over time.