ICA_group3

group 3

2024-04-07

Import data

```
substance_use <- read.csv("substance_use.csv")
head(substance_use)
```

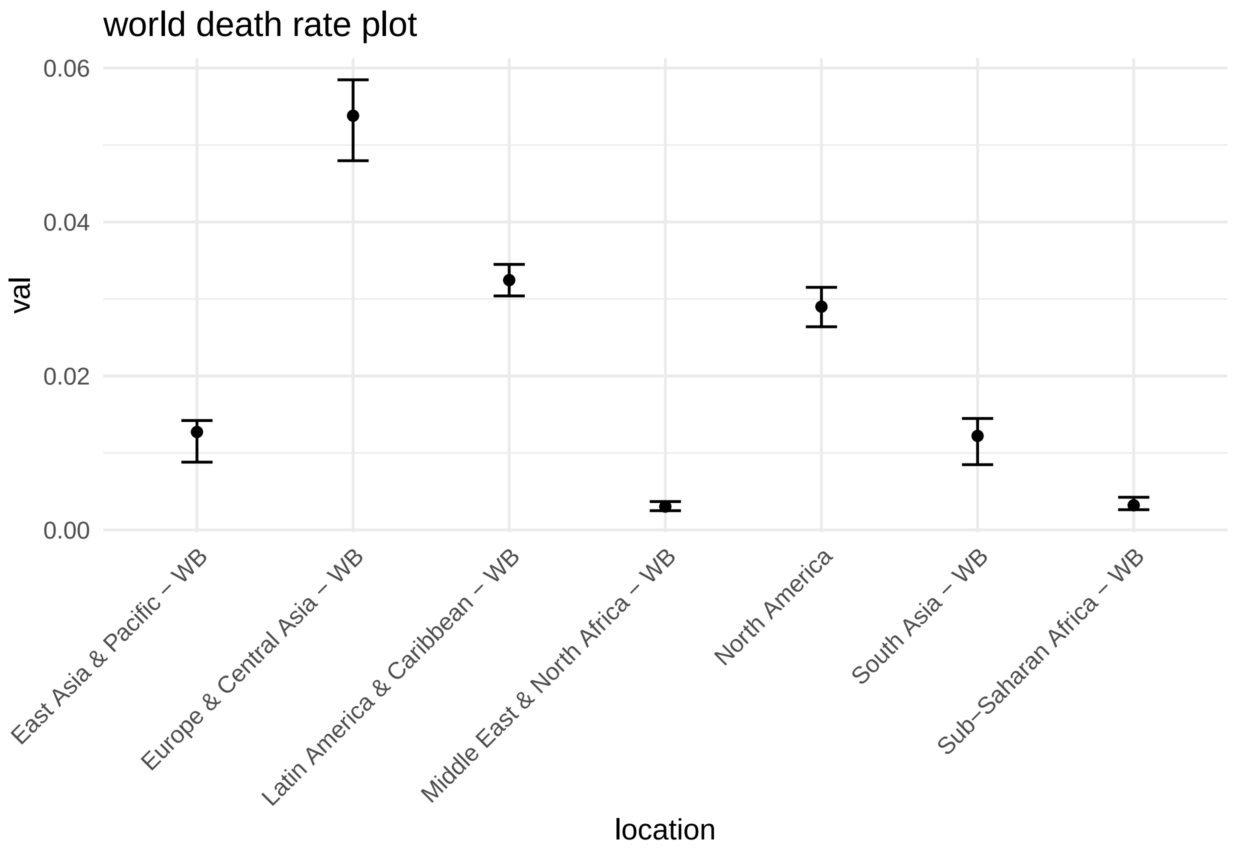
```
##   measure                  location    sex      age                cause
## 1  Deaths East Asia & Pacific - WB   Male 25 to 29 Alcohol use disorders
## 2  Deaths East Asia & Pacific - WB Female 25 to 29 Alcohol use disorders
## 3  Deaths East Asia & Pacific - WB   Male 30 to 34 Alcohol use disorders
## 4  Deaths East Asia & Pacific - WB Female 30 to 34 Alcohol use disorders
## 5  Deaths East Asia & Pacific - WB   Male 35 to 39 Alcohol use disorders
## 6  Deaths East Asia & Pacific - WB Female 35 to 39 Alcohol use disorders
##    metric year         val       upper       lower
## 1 Percent 1990 0.004355489 0.005574785 0.003579575
## 2 Percent 1990 0.002316023 0.002622133 0.002052042
## 3 Percent 1990 0.006539015 0.007974114 0.005392593
## 4 Percent 1990 0.002667792 0.002950154 0.002417720
## 5 Percent 1990 0.007597508 0.010585770 0.006359210
## 6 Percent 1990 0.002744876 0.003049935 0.002468063
```

1. In 2019, what region of the world has the highest rate of alcohol-related deaths among men aged 40-44?

```
# Select the data
Q1 <- subset(substance_use,year=="2019" & sex=="Male" & age=="40 to 44"
             & cause=="Alcohol use disorders" & measure=="Deaths")

# Choose the region with the highest val
highest_region <- (Q1[which.max(Q1$val), ])$location
print(highest_region)
```

```
## [1] "Europe & Central Asia - WB"
```

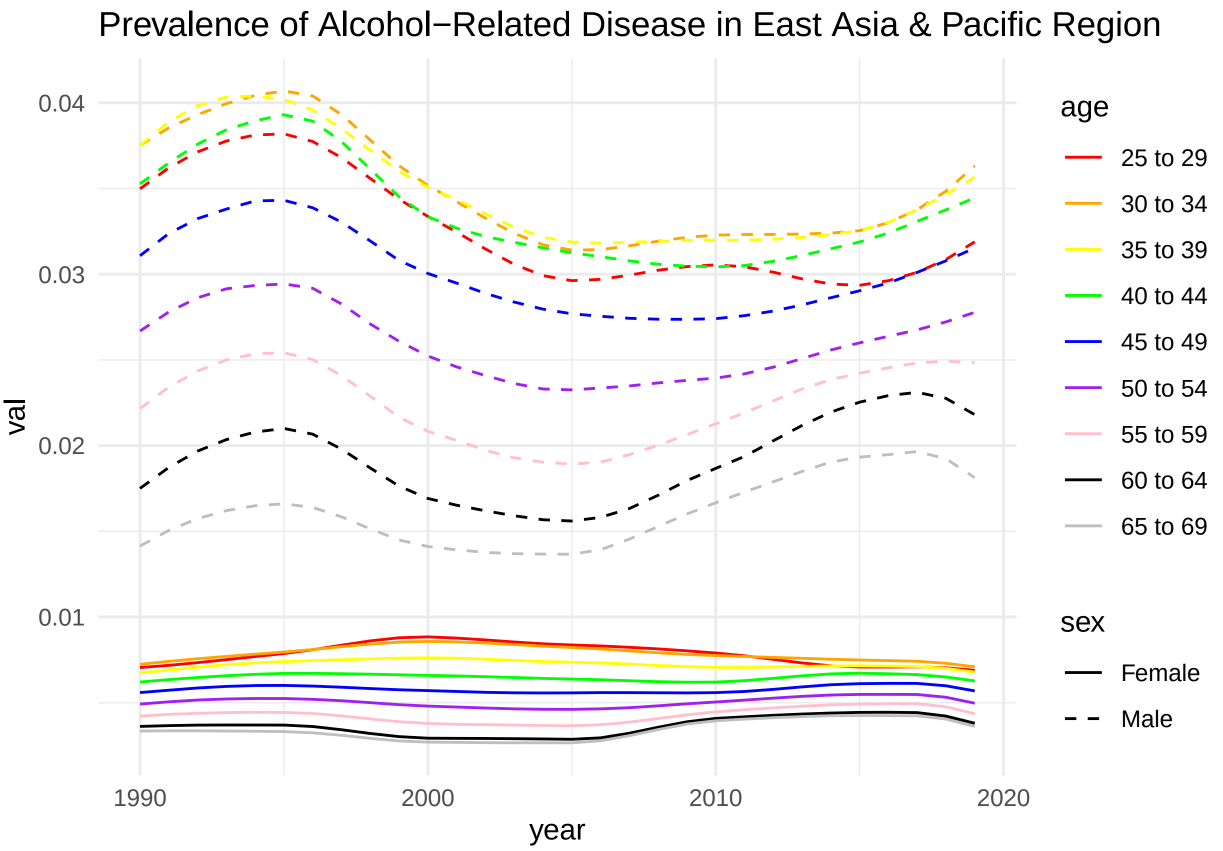
```
# Draw the death rate plot
ggplot(Q1, aes(x = location, y = val, ymin = lower, ymax = upper)) +
  geom_point() +
  geom_errorbar(width = 0.2, position = position_dodge(0.9)) +
  labs(x = "location", y = "val", title = 'world death rate plot') +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



1. **Conclusion: Based on the max function and the plot, we can observe the region with the highest rate of alcohol-related deaths is "Europe & Central Asia - WB".**

2.1 Looking at the prevalence of alcohol-related disease in the East Asia and Pacific region, how has this changed over time and in the different age groups?

```
# Select the East Asia and Pacific region
Q2 <- subset(substance_use,cause=="Alcohol use disorders" & measure=="Prevalence"
             & location=="East Asia & Pacific - WB")

# Plot the prevalence among the different time and age groups
ggplot(Q2, aes(x = year, y = val, color = age, linetype = sex)) +
  geom_line() +
  labs(x = "year", y = "val", color = "age", linetype = "sex",
       title = 'Prevalence of Alcohol-Related Disease in East Asia & Pacific Region') +
  scale_color_manual(values = c("red", "orange", "yellow", "green","blue",
                                "purple", "pink", "black", "gray")) +
  scale_linetype_manual(values = c("solid", "dashed")) +
  theme_minimal()
```



2.1 Conclusion: There was no significant change in the trend for females. For males, the prevalence increases between 1990-1995, decreases between 1995-2005, and increases again between 2005-2019. The prevalence is higher in males than in females.

2.2 Is there a difference between men and women?

H0: There is no difference between men and women. HA: There is a difference between men and women.

```r
# Select the sex-related prevalence data
opioid_gender <- Q2[,c("year", "val", "age","sex")]
p_value<- c()
age.l <- unique(opioid_gender$age) # Select the different year
for (i in age.l){
  male <- filter(opioid_gender, sex == "Male", age == i)
  male <- male[order(male$year),]
  female <- filter(opioid_gender, sex == "Female", age == i)
  female <- female[order(female$year),]
  result <- wilcox.test(male$val, female$val, paired = TRUE, alternative = "two.sided")
  p_value <- c(p_value, result$p.value) # Store the results from wilcox.test
}
print(p_value)
```

```
## [1] 1.862645e-09 1.862645e-09 1.862645e-09 1.862645e-09 1.862645e-09
## [6] 1.862645e-09 1.862645e-09 1.862645e-09 1.862645e-09
```

```r
print(p_value < 0.05) # Display the compare results
```

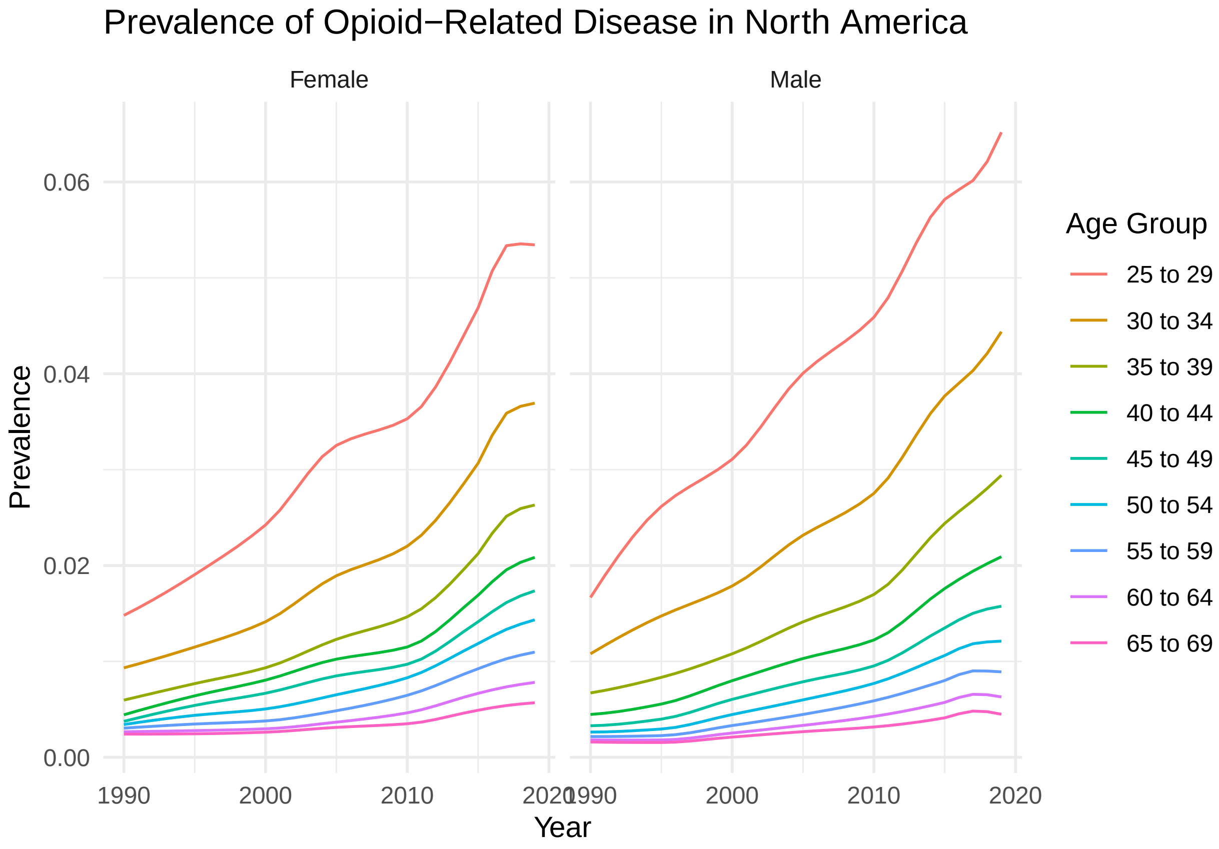
```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

2.2 Conclusion: The men and women in different years were paired for wilcox.test, and the P-values were all less than 0.05, so H0 was rejected, There is a difference between men and women.

3. In the United States, there is talk of an "Opioid epidemic". Part of the problem is that since the late 1990s, doctors have increasingly been prescribing pain killers which can be highly addictive.

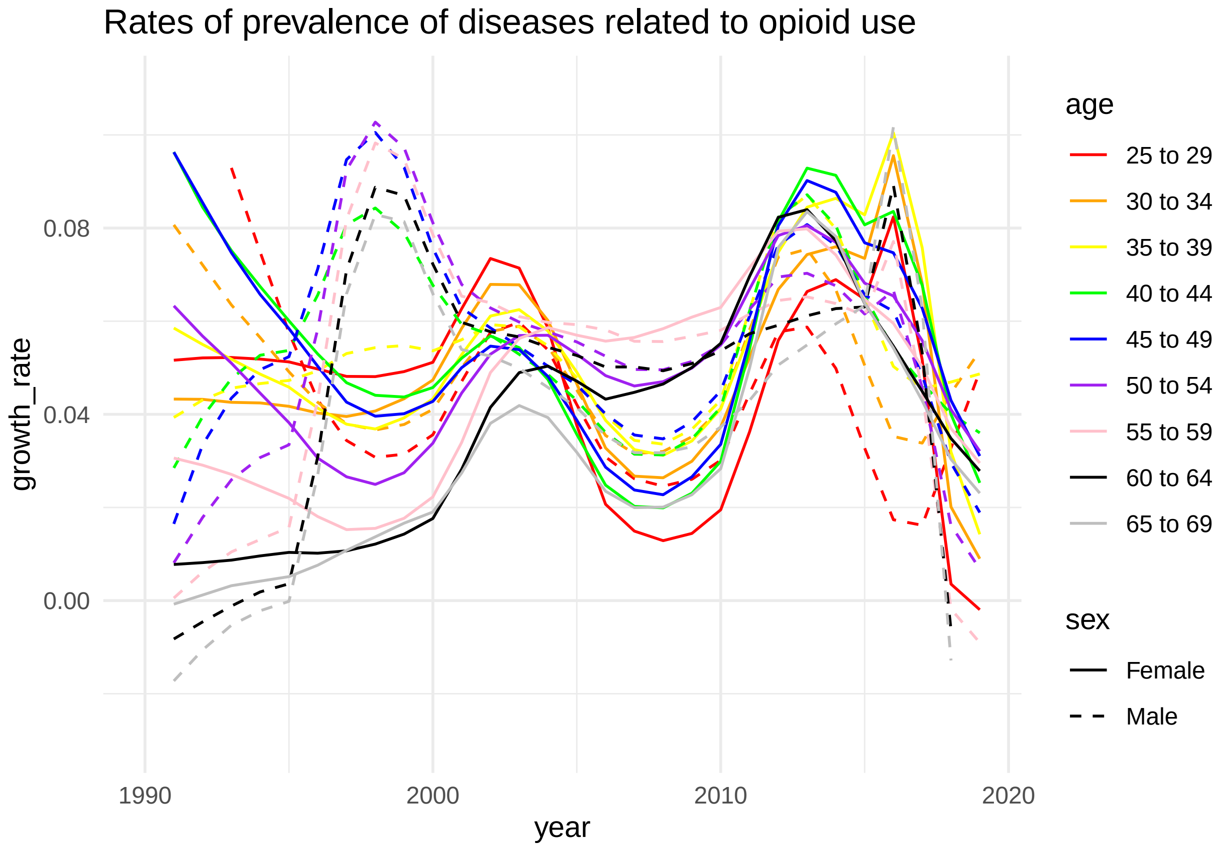
3.1 Looking at the data from the United States, can you confirm an increase in the prevalence of diseases related to opioid use?

```r
# Select the data
Q3 <- substance_use %>%
    filter(location == "North America", cause == "Opioid use disorders",
           measure == "Prevalence")

# Plot the prevalence of opioid related diseases
ggplot(Q3, aes(x = year, y = val, color = age)) +
  geom_line() +
  facet_wrap(~sex) +
  labs(title = "Prevalence of Opioid-Related Disease in North America",
       x = "Year",
       y = "Prevalence",
       color = "Age Group") +
  theme_minimal()
```



```r
# Compare the next year's data with the previous year's, calculate the growth rate.
Q3 <- Q3[order(Q3$sex, Q3$age, Q3$year), ]
Q3$growth_rate <- c(NA, (Q3$val[-1] / Q3$val[-nrow(Q3)])-1)
Q3$growth_rate[!duplicated(Q3$age)] <- NA

# Polt the growth rate in different years
ggplot(Q3, aes(x = year, y = growth_rate, color = age, linetype = sex)) +
  geom_line() +scale_y_continuous(limits = c(-0.03,0.11))+
  labs(x = "year", y = "growth_rate", color = "age", linetype = "sex",
       title = 'Rates of prevalence of diseases related to opioid use') +
  scale_color_manual(values = c("red", "orange", "yellow", "green","blue",
                                "purple", "pink", "black", "gray")) +
  scale_linetype_manual(values = c("solid", "dashed")) +
  theme_minimal()
```

```
## Warning: Removed 22 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```



Based on the graph, the growth rate of the prevalence of Opioid-Related Disease among males reached its peak in 1997. Notably, we observed that the highest point of the growth rate for all age groups of females is generally delayed by five years compared to males. Here, we infer that there is a delayed effect for females in developing Opioid-Related Disease after taking opioid medications compared to males. Based on the situation where doctors have been overprescribing painkillers since the late 1990s, we deduce that the starting point of the increasingly use of opioid was in 1997.

```r
# Subset the origin data
filtered_data_before <- substance_use %>%
  filter(location == "North America", cause == "Opioid use disorders",
         measure == "Prevalence", year %in% 1990:1997)
filtered_data_after <- substance_use %>%
  filter(location == "North America", cause == "Opioid use disorders",
         measure == "Prevalence", year %in% 1997:2019)

# Set a data frame for storage
male_slopes_before <- data.frame(age = character(), slope = numeric())
female_slopes_before <- data.frame(age = character(), slope = numeric())

# Perform linear regression analysis
for (i in unique(filtered_data_before$age)) {
  age_data = filtered_data_before[filtered_data_before$age == i,]
  male_lm = lm(val~year,data=age_data[age_data$sex=="Male",])
  female_lm = lm(val~year,data=age_data[age_data$sex=="Female",])
  # Store the slopes and years
  male_slopes_before <- rbind(male_slopes_before,
                              data.frame(age = i, slope = coef(male_lm)[2]))
  female_slopes_before <- rbind(female_slopes_before,
                                data.frame(age = i, slope = coef(female_lm)[2]))
}

# Set a data frame for storage
male_slopes_after <- data.frame(age = character(), slope = numeric())
female_slopes_after <- data.frame(age = character(), slope = numeric())

# Perform Linear regression analysis
for (i in unique(filtered_data_after$age)) {
  age_data = filtered_data_after[filtered_data_after$age == i,]
  male_lm = lm(val~year,data=age_data[age_data$sex=="Male",])
  female_lm = lm(val~year,data=age_data[age_data$sex=="Female",])
  # Store the slopes and years
  male_slopes_after <- rbind(male_slopes_after,
                             data.frame(age = i, slope = coef(male_lm)[2]))
  female_slopes_after <- rbind(female_slopes_after,
                               data.frame(age = i, slope = coef(female_lm)[2]))
}

# Test the slope for male
male_test_result <- wilcox.test(male_slopes_after$slope, male_slopes_before$slope,
                                alternative = "greater", paired = TRUE)
```

```r
# Test the slope for female
female_test_result <- wilcox.test(female_slopes_after$slope, female_slopes_before$slope,
                                  alternative = "greater", paired = TRUE)

# Print the test results
print(male_test_result)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  male_slopes_after$slope and male_slopes_before$slope
## V = 45, p-value = 0.001953
## alternative hypothesis: true location shift is greater than 0
```

```r
print(female_test_result)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  female_slopes_after$slope and female_slopes_before$slope
## V = 45, p-value = 0.001953
## alternative hypothesis: true location shift is greater than 0
```

3.1 Conclusion: Using linear regression, we compare the slopes of males and females prevalence rates before and after 1997. From the results of one-tail (greater) Wilcox.test, the p-values for males and females are both less than 0.05. So we can confirm an increase in the prevalence of diseases related to opioid use since the late 1990s.

3.2 What age group is the most affected?

```r
male_slopes_before$sex <- "Male"
female_slopes_before$sex <- "Female"
male_slopes_after$sex <- "Male"
female_slopes_after$sex <- "Female"

# Combine slope data for male and female
slopes_before <- rbind(male_slopes_before, female_slopes_before)
slopes_after <- rbind(male_slopes_after, female_slopes_after)

rownames(slopes_before) <- NULL
rownames(slopes_after) <- NULL

# Calculate the slope difference for each age group
slope_diff <- merge(slopes_after, slopes_before, by = c("age", "sex"),
                    suffixes = c("_after", "_before"))
slope_diff$slope_diff <- slope_diff$slope_after - slope_diff$slope_before

# Find the age group with the largest slope difference for males
```

```r
max_diff_age_male <- slope_diff[slope_diff$sex == "Male",][
  which.max(slope_diff[slope_diff$sex == "Male",]$slope_diff), "age"]

# Find the age group with the largest slope difference for females
max_diff_age_female <-
  slope_diff[slope_diff$sex == "Female",][
    which.max(slope_diff[slope_diff$sex == "Female",]$slope_diff), "age"]

# Print the slope difference results
print(max_diff_age_male)
```

```
## [1] "30 to 34"
```

```r
print(max_diff_age_female)
```

```
## [1] "30 to 34"
```

3.2 Conclusion: We calculate the differences in slopes for males and females in different age groups before and after 1997. From the results, the 30-34 age groups in both male and female show the most significant change in slope after 1997, so we confirm that the 30-34 age group is the most affected. Also, based on the 'Prevalence of Opioid-Related Disease in North America' plot, we find the 25 to 29 age groups both in male and female have the highest prevalence of diseases related to opioid use.

4. **Ask your own question**

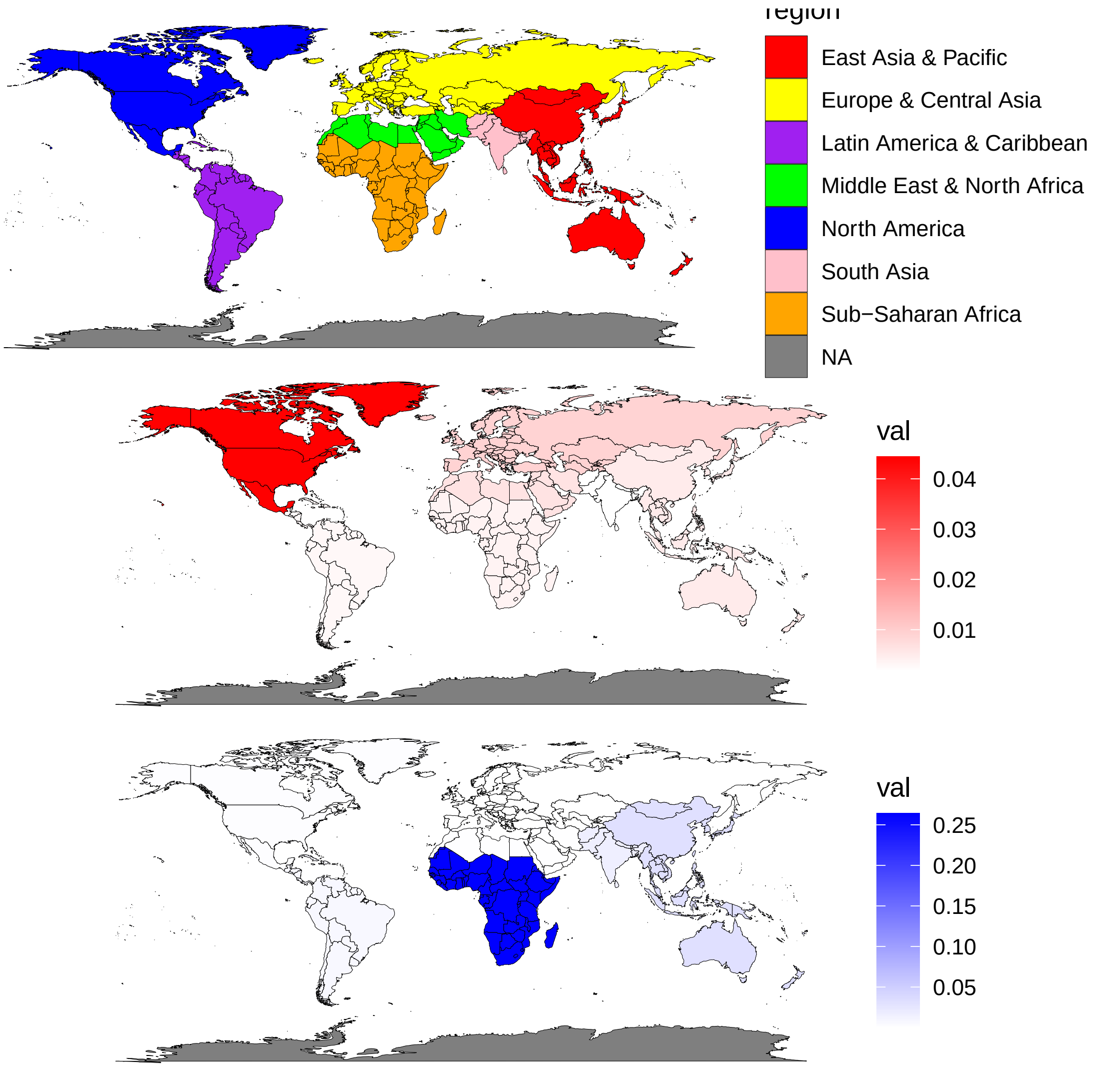
The 30-34 age group is known to be most affected in the prevalence of diseases related to opioid use since the late 1990s. We want to know how the prevalence and death rates related to the opioid use in this age group are distributed across different countries and regions. So we choose year 2019 to see how the prevalence and deaths of male are distributed across the world. For method, we use ggplot2 geom_map to visulize the data.

```
all_countries <- c(east_asia_pacific, north_america, middle_east_north_africa,
                   europe_central_asia, latin_america_caribbean, sub_saharan_africa, south_asia)
all_regions <- c(rep("East Asia & Pacific", length(east_asia_pacific)),
                 rep("North America", length(north_america)),
                 rep("Middle East & North Africa", length(middle_east_north_africa)),
                 rep("Europe & Central Asia", length(europe_central_asia)),
                 rep("Latin America & Caribbean", length(latin_america_caribbean)),
                 rep("Sub-Saharan Africa", length(sub_saharan_africa)),
                 rep("South Asia", length(south_asia)))

df <- data.frame(country = all_countries, region = all_regions)

df_region = unique(df$region)

# Get world map data
world_map <- map_data("world")

# Merge map data with your data box
world_map$region <- df$region[match(world_map$region, df$country)]

# Create a color map
color_mapping_region <- c("East Asia & Pacific" = "red",
                   "North America" = "blue",
                   "Middle East & North Africa" = "green",
                   "Europe & Central Asia" = "yellow",
                   "Latin America & Caribbean" = "purple",
                   "Sub-Saharan Africa" = "orange",
                   "South Asia" = "pink")

# Draw map
g1 <- ggplot() +
  geom_polygon(data = world_map,
               aes(x = long, y = lat, group = group, fill = region),
               color = "black",
               size = 0.1) +
  scale_fill_manual(values = color_mapping_region) +
  coord_fixed() +
  theme_void()
```

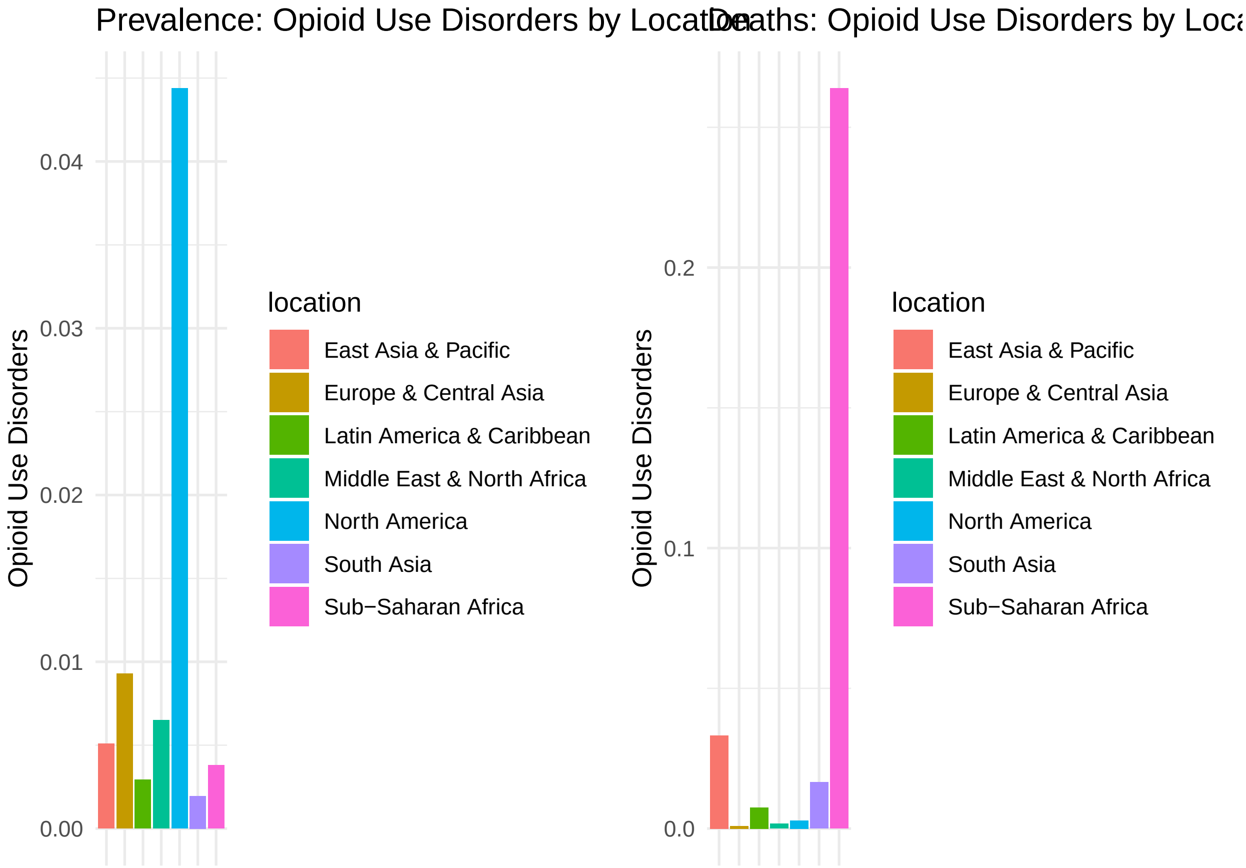
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
# Suppose you have a data box with country and corresponding disease data
opioid_data <- substance_use %>%
  filter(year == 2019, sex == "Male", age == "30 to 34",
         cause == "Opioid use disorders", measure == "Prevalence")

opioid_data$location = c("North America", "South Asia",
                         "Latin America & Caribbean","Sub-Saharan Africa",
                         "Middle East & North Africa" ,"East Asia & Pacific",
                         "Europe & Central Asia")

# Merge disease data with map data
world_map$val <- opioid_data$val[match(world_map$region, opioid_data$location)]

# Create a color map
color_mapping_prevalence <- colorRampPalette(c("white", "red"))

# Draw prevalence map
g2<- ggplot() +
  geom_polygon(data = world_map,
               aes(x = long, y = lat, group = group, fill = val),
               color = "black",
               size = 0.1) +
  scale_fill_gradientn(colors = color_mapping_prevalence(100)) +
  coord_fixed() +
  theme_void()

# Suppose you have a data box with country and corresponding disease data
opioid_data_deaths <- substance_use %>%
  filter(year == 2019, sex == "Male", age == "30 to 34",
         cause == "Opioid use disorders", measure == "Deaths")

opioid_data_deaths$location = c("North America", "South Asia",
                         "Latin America & Caribbean","Sub-Saharan Africa",
                         "Middle East & North Africa" ,"East Asia & Pacific",
                         "Europe & Central Asia")

# Merge disease data with map data
world_map$val <- opioid_data_deaths$val[match(world_map$region, opioid_data_deaths$location)]

# Create a color map
color_mapping_death <- colorRampPalette(c("white", "blue"))

# Draw death map
g3 <- ggplot() +
  geom_polygon(data = world_map,
               aes(x = long, y = lat, group = group, fill = val),
               color = "black",
               size = 0.1) +
  scale_fill_gradientn(colors = color_mapping_death(100)) +
  coord_fixed() +
  theme_void()
```

```
# Print map
grid.arrange(g1, g2, g3, nrow = 3)
```



```
# Use the opioid data data box to draw a bar chart
p1 <- ggplot(opioid_data, aes(x = location, y = val, fill = location)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(y = "Opioid Use Disorders", title = "Prevalence: Opioid Use Disorders by Location") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.x = element_blank())
p2 <- ggplot(opioid_data_deaths, aes(x = location, y = val, fill = location)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(y = "Opioid Use Disorders", title = "Deaths: Opioid Use Disorders by Location") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.x = element_blank())
grid.arrange(p1, p2, ncol = 2)
```



4. Conclusion: From the prevalence map, we can see that the North America is most affected. North America, especially the USA, has a significantly higher rate of opioid prevalence. After searching for information, we found some reasons for it. 1)The medical system: Doctors in North America are more likely to give prescriptions with opioid medicine to deal with pain. 2) The medicine industry has great influence in North America. Their promotion and marketing, along with their false advertising regarding the safety and addictiveness of opioid medications, have facilitated the over-prescription and excessive use of these drugs. 3) Poverty, unemployment, and a lack of social support networks increase the risk of drug abuse. 4) Although regulation has tightened in recent years, prior to the epidemic, the oversight of opioid medications was relatively lax, making these drugs easily susceptible to abuse. Considering the reasons mentioned above, opioids are more prevalent in North America than in other regions.

From the deaths map, we can see that the Sub-Sahara Africa is most affected. Although the use of opioids in sub-Saharan Africa is not high, due to the low level of medical care, once prevalence, it is difficult to treat and face the risk of death.