ADS2 Group Exercise ICA

Group 6

2024-04-02

Part 0 : Clean and tidy the dataset

```
library(ggplot2)
library(dplyr)
library(tidyverse)
library(tidyr)
library(knitr)

substance_use = read.csv("substance_use.csv")
```

Any missing values?

```
anyNA(substance_use)
```

```
## [1] FALSE
```

```
sum(substance_use[ , ] == "")
```

```
## [1] 0
```

There is no missing values in the dataset.

Any duplicates?

```
which(duplicated(substance_use))
```

```
## integer(0)
```

There is no duplicate in the dataset.

Typos, naming schemes and data types?

```
str(substance_use)
```

```
## 'data.frame':    15120 obs. of  10 variables:
##  $ measure : chr  "Deaths" "Deaths" "Deaths" "Deaths" ...
##  $ location: chr  "East Asia & Pacific - WB" "East Asia & Pacific - WB" "East Asia & Pacific - WB" "
```

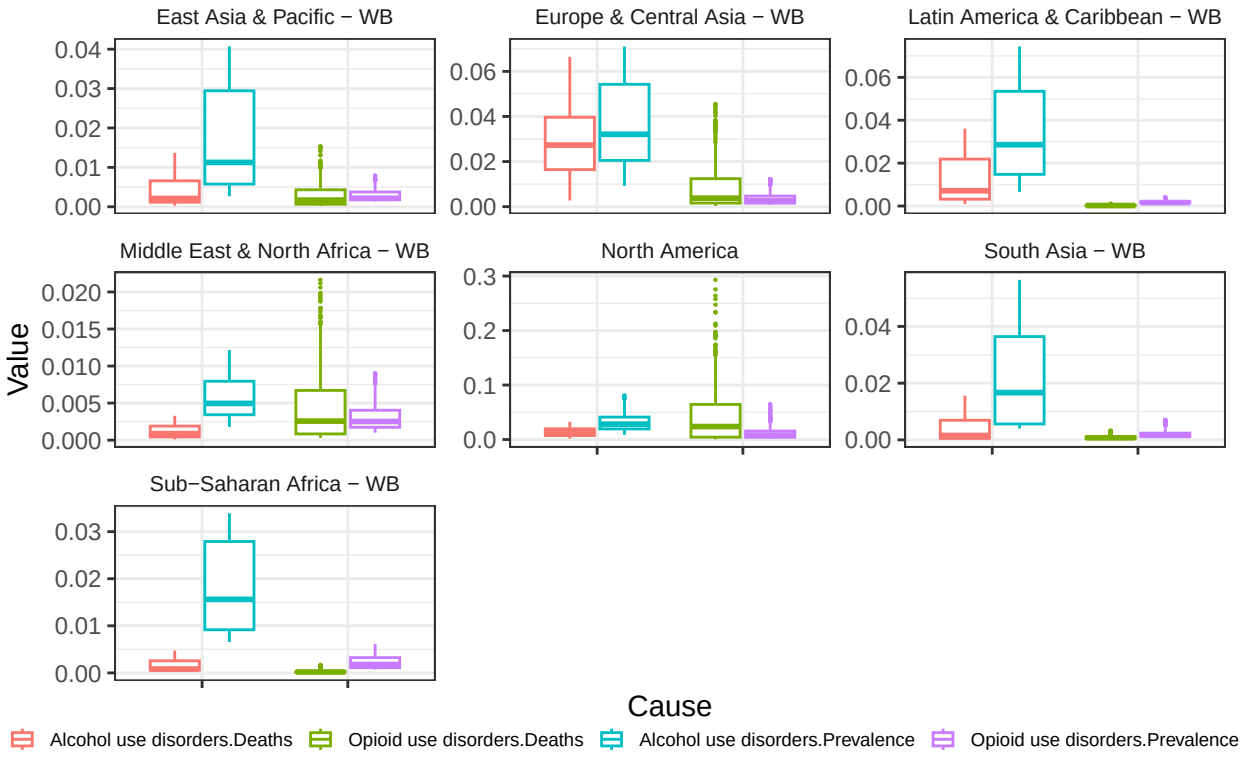
```
##  $ sex     : chr  "Male" "Female" "Male" "Female" ...
##  $ age     : chr  "25 to 29" "25 to 29" "30 to 34" "30 to 34" ...
##  $ cause   : chr  "Alcohol use disorders" "Alcohol use disorders" "Alcohol use disorders" "Alcohol us
##  $ metric  : chr  "Percent" "Percent" "Percent" "Percent" ...
##  $ year    : int  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
##  $ val     : num  0.00436 0.00232 0.00654 0.00267 0.0076 ...
##  $ upper   : num  0.00557 0.00262 0.00797 0.00295 0.01059 ...
##  $ lower   : num  0.00358 0.00205 0.00539 0.00242 0.00636 ...
```

The location, sex and age variables can be factors.

```
substance_use$location = as.factor(substance_use$location)
substance_use$sex = as.factor(substance_use$sex)
substance_use$age = as.factor(substance_use$age)
```

Any outliers or strange patterns?

```
ggplot(data = substance_use,
       aes(x = cause, y = val, color = interaction(cause, measure))) +
  geom_boxplot(outlier.size = 0.1) +
  facet_wrap(~location, scales = "free_y") +
  labs(x = "Cause", y = "Value", color = "") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        strip.text = element_text(size = 8),
        strip.background =  element_rect(fill = NA, colour = NA),
        legend.position = "bottom",
        legend.key.size = unit(10, "pt"),
        legend.text = element_text(size = 7),
        legend.margin = margin(-10, 0, 0, -40))
```




```
which(substance_use$val < 0)
```

```
## integer(0)
```

```
which(substance_use$val > 1)
```

```
## integer(0)
```

Figure 0: The boxplot of the death rate and prevalence of the alcohol and opioid use disorders in different locations.

The range of the value is normal. From Figure 0, You can see some outliers. However, if you draw the plots for each location, age and sex, you will find that the data looks pretty nice. Due to the page limit, we do not show the plots. Therefore, we decide not to delete the outliers in Figure 0.

Part 1: Exploring the data

In this part, we will answer the questions given in the guidance.

Question 1

Description

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

Method

We extract the data of the alcohol-related death rate among men aged 40-44 and sort the location according to the death rate in the descending order.

```
highest_alcohol_deaths = substance_use %>%
  filter(measure == "Deaths", year == 2019, age == "40 to 44",
         sex == "Male", cause == "Alcohol use disorders") %>%
  select(location, val) %>%
  arrange(desc(val)) %>%
  top_n(1, val)

kable(highest_alcohol_deaths)
```

location	val
Europe & Central Asia - WB	0.0537985

Conclusion

In 2019, Europe & Central Asia has the highest rate of alcohol-related deaths among men aged 40-44.

Question 2

Description

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? Is there a difference between men and women?

Method

First, we extract the data.

```
eap_alcohol_data = substance_use %>%
  filter(measure == "Prevalence", cause == "Alcohol use disorders",
         location == "East Asia & Pacific - WB") %>%
  select(year, age, sex, val)
```

Next, we visualize the data in two ways, faceting the figures by age groups and by sex groups.

```
eap_alcohol_data1 = eap_alcohol_data
eap_alcohol_data1$age = paste0(eap_alcohol_data1$age, " years old")
ggplot(eap_alcohol_data1,
       aes(x = year, y = val, color = sex,
           group = interaction(sex, age))) +
  geom_line() +
  facet_wrap(~age, scales = 'free_y') +
  labs(x = "Year", y = "Prevalence (%)", color = "Sex") +
  theme_bw() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5),
        strip.background =  element_rect(fill = NA, colour = NA))
```

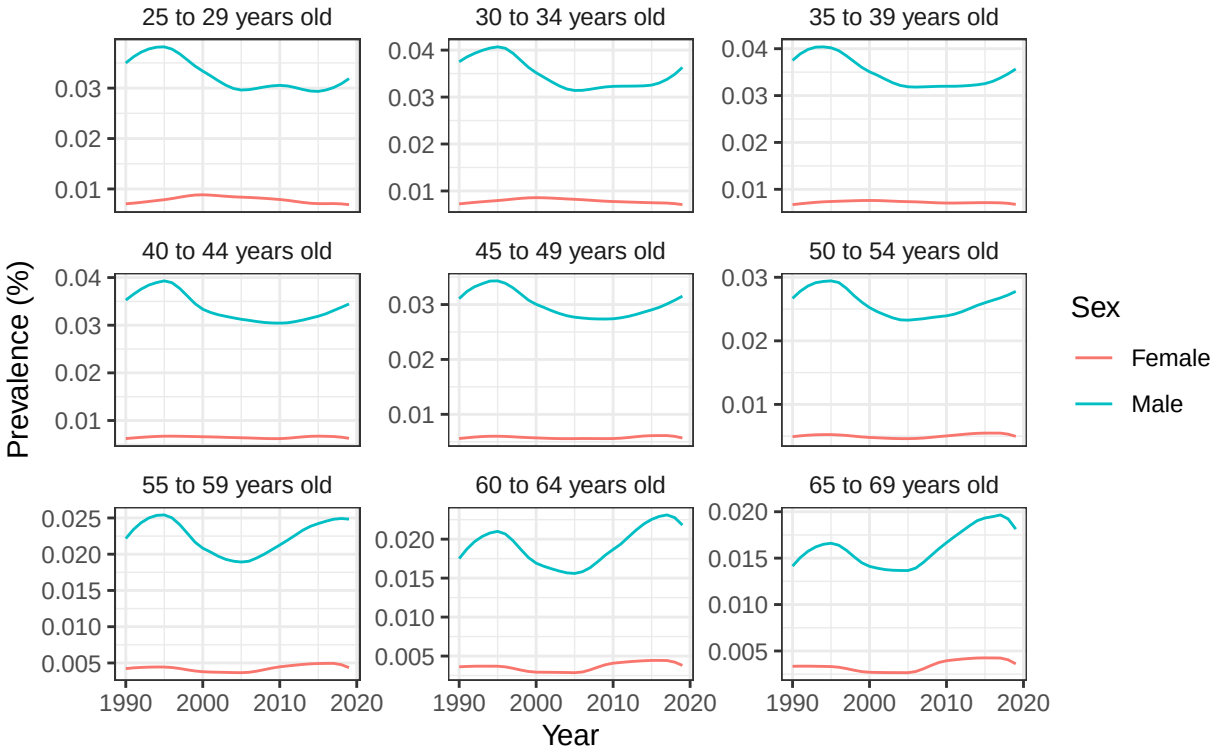


Figure 2.1: The trends of the alcohol-related disease prevalence in east Asia and Pacific. This figure is faceted by age groups, better for visually comparing the prevalence between the male and the female.

```
ggplot(eap_alcohol_data,
       aes(x = year, y = val, color = age,
           group = interaction(sex, age))) +
  geom_line() +
  geom_point(size = 0.6) +
  facet_wrap(~sex, scales = 'free_y') +
  labs(x = "Year", y = "Prevalence (%)", color = "Age (years old)") +
  theme_bw() +
  theme(legend.title = element_text(hjust = 0.5),
        plot.title = element_text(hjust = 0.5),
        strip.background =  element_rect(fill = NA, colour = NA))
```

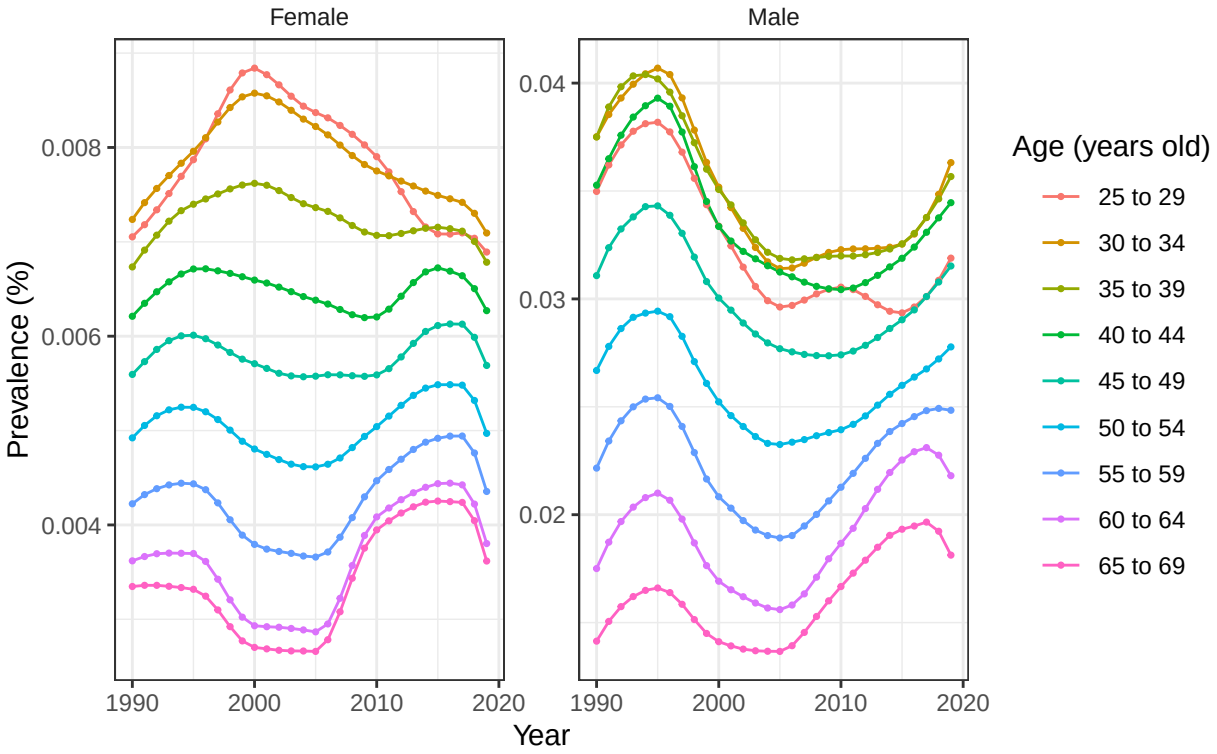


Figure 2.2: The trends of the alcohol-related disease prevalence in east Asia and Pacific. This figure is faceted by sex groups, better for visually comparing the prevalence between different age groups.

From Figure 2.1 and Figure 2.2, you can see how the prevalence of the alcohol-related disease in the East Asia and Pacific region changes over time in different age groups. Also, you can identify different patterns of the prevalence between male and female.

In the following, we will validate the difference between male and female in a more statistical way.

First, we formulate the hypotheses.

- The null hypothesis (H0): There is no difference in the prevalence of different age groups between male and female.
- The alternative hypothesis (HA): There is difference in the prevalence of different age groups between male and female.

The data for male and female in the same year paired, and we want to test the differences in prevalence between male and female, so we will use the paired statistical test in the following. First, We test the normality of the differences in prevalence.

- If the differences are normally distributed, we will perform the parametric Student's t-test to test the hypotheses.
- If the differences are not normally distributed, we will perform the non-parametric Wilcoxon rank sum test to test the hypotheses.

```
eap_alcohol_data2 = eap_alcohol_data %>%
  group_by(age) %>%
  group_split() %>%
  map(~spread(., key = "sex", value = "val")) %>%
```

```
  map(~{
    shapiro_p = shapiro.test(.$Male - .$Female)$p.value
    if(shapiro_p < 0.05){
      test_type = "Wilcoxon rank sum test"
      p_value = wilcox.test(.$Male, .$Female, paired = T)$p.value
    }
    else{
      test_type = "Student's t-test"
      p_value = t.test(.$Male, .$Female, paired = T)$p.value
    }
    tibble(shapiro_p = shapiro_p, test_type = test_type, p_value = p_value)
  }) %>%
  bind_rows() %>%
  mutate(age = unique(eap_alcohol_data$age), .before = shapiro_p)
kable(eap_alcohol_data2, digit = 12)
```

age	shapiro_p	test_type	p_value
25 to 29	0.0001992235	Wilcoxon rank sum test	1.863e-09
30 to 34	0.0012807353	Wilcoxon rank sum test	1.863e-09
35 to 39	0.0002517598	Wilcoxon rank sum test	1.863e-09
40 to 44	0.0006353554	Wilcoxon rank sum test	1.863e-09
45 to 49	0.0023908823	Wilcoxon rank sum test	1.863e-09
50 to 54	0.0044001598	Wilcoxon rank sum test	1.863e-09
55 to 59	0.0288893147	Wilcoxon rank sum test	1.863e-09
60 to 64	0.0411236218	Wilcoxon rank sum test	1.863e-09
65 to 69	0.0185078808	Wilcoxon rank sum test	1.863e-09

Conclusion

Of all the age groups, the p values from the statistical test are all less than 10^{-8}. Therefore, we can reject the null hypothesis (H0). There is sufficient evidence to conclude that there is significant difference in the prevalence between the male and the female within each age group.

Question 3

Description

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. Looking at the data from the United States, can you confirm an increase in the prevalence of diseases related to opioid use? What age group is the most affected?

Method

There is no data named "United States" in the locations divided by the world bank, so we extract the data from "North America" to represent the data from the United States.

According to the question, we will separate the data from 1990 to 2019 into two parts "before the start of increasing prescription of the opioid" (1990-1997) and "after the start of increasing prescription of the opioid" (1998-2019). We will abbreviate the two periods to the "before" period and the "after" period in the following.

```
opioid_use_na = substance_use %>%
  filter(measure == "Prevalence",
         location == "North America",
         cause == "Opioid use disorders") %>%
  select(age, year, sex, val)
```

Next, we visualize the data in two ways, faceting the figures by age groups and by sex groups.

```
opioid_use_na1 = opioid_use_na
opioid_use_na1$age = paste0(opioid_use_na1$age, " years old")
ggplot(opioid_use_na1,
       aes(x = year, y = val, color = sex,
           group = interaction(sex, age))) +
  geom_line() +
  geom_vline(xintercept = 1998, color = "darkgray") +
  facet_wrap(~age, scales = 'free_y') +
  labs(x = "Year", y = "Prevalence (%)", color = "Sex") +
  theme_bw() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45),
        strip.background =  element_rect(fill = NA, colour = NA))
```

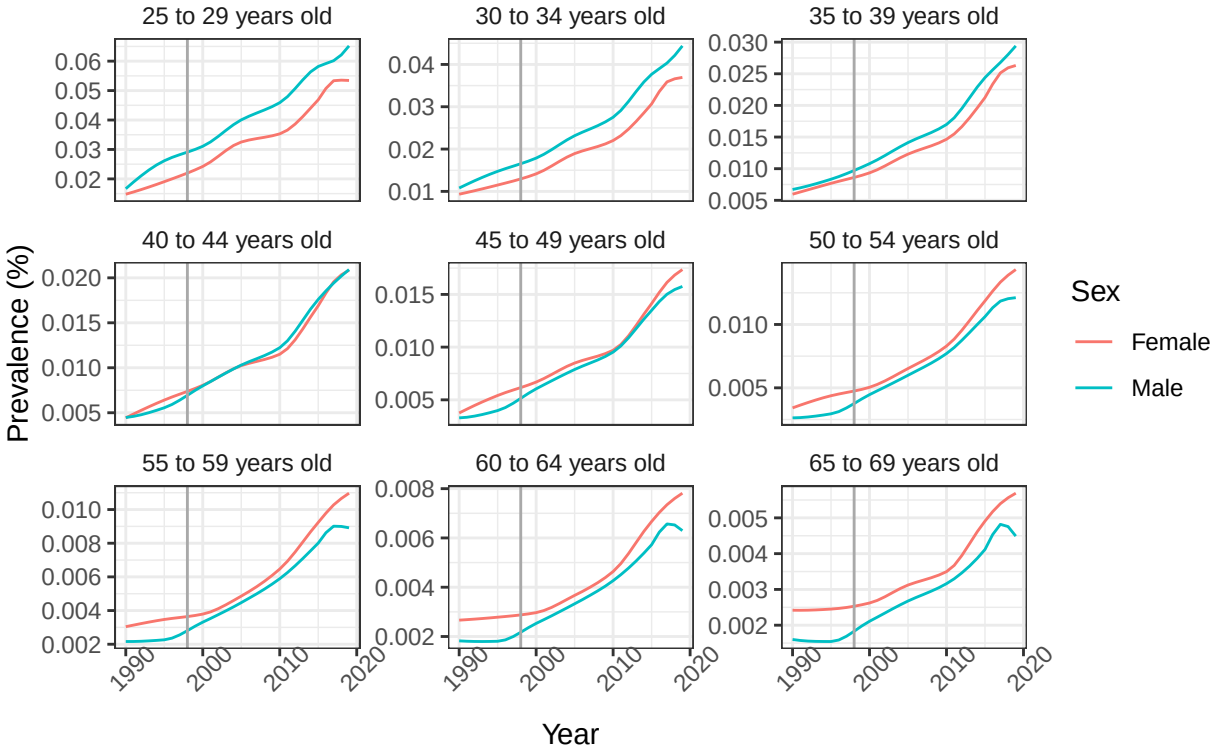


Figure 3.1: The trends of the opioid use related disease prevalence in North America. The figure is faceted by age groups, better for visually comparing the prevalence between male and female.

```r
ggplot(opioid_use_na,
       aes(x = year, y = val, color = age,
           group = interaction(sex, age))) +
  geom_line() +
  geom_point(size = 0.7) +
  geom_vline(xintercept = 1998, color = "darkgrey") +
  facet_wrap(~sex, scales = 'free_y') +
  labs(x = "Year", y = "Prevalence (%)", color = "Age (years old)") +
  theme_bw() +
  theme(legend.title = element_text(hjust = 0.5),
        plot.title = element_text(hjust = 0.5),
        strip.background =  element_rect(fill = NA, colour = NA))
```

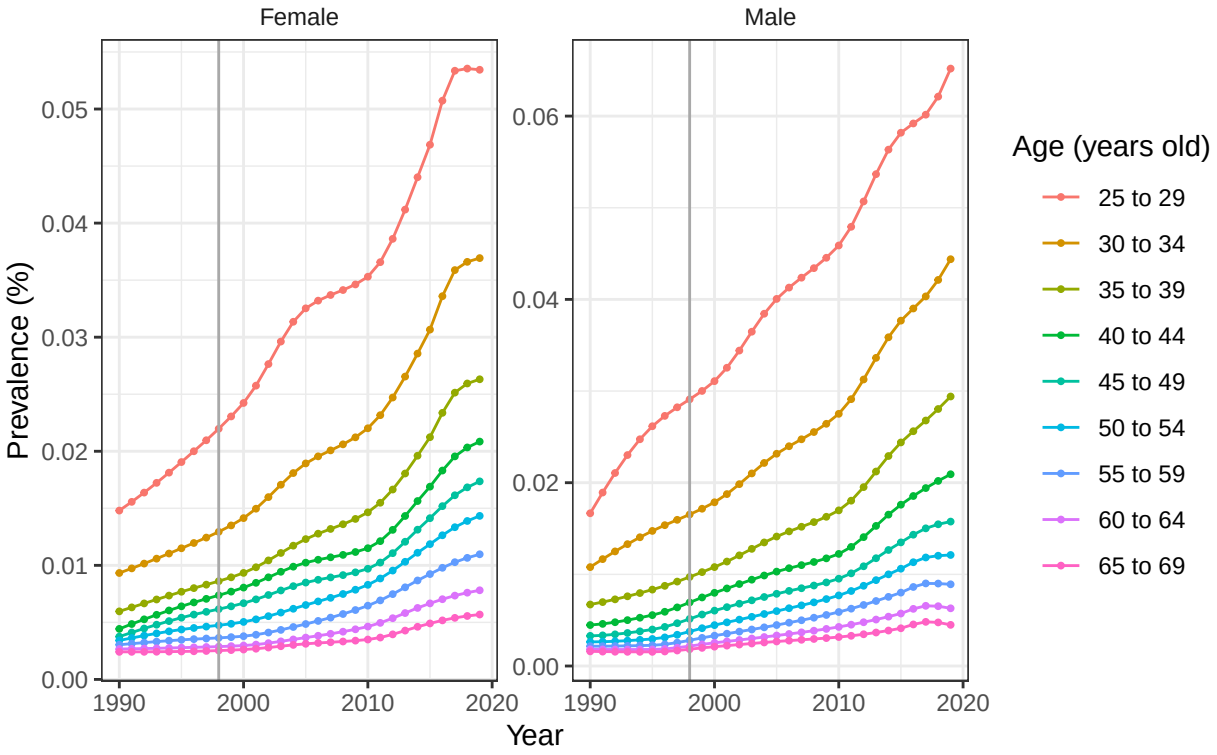


Figure 3.2: The trends of the opioid use related disease prevalence in North America. The figure is faceted by sex groups, better for visually comparing the prevalence between different age groups.

From Figure 3.1 and Figure 3.2, we can identify that there is an overall increase pattern within each age and sex group. Since the prevalence of the "before" and "after" period are both increasing, and we want to confirm whether the increasing prescription the opioid has an effect on the prevalence, we should compare the speed (slope) of the increase in the prevalence rather than the prevalence itself. If the prevalence of the "after" period is increasing faster than the prevalence of the "before" period, the increasing prescription of the opioid may have an effect.

To test whether the slope of the prevalence of the "before" period is different from the slope of the prevalence of the "after" period, we formulate the hypotheses.

- The null hypothesis (H0): There is no difference in the slope of the prevalence between the "before" period and the "after" period.

sex	age	p_value	effect_size	sex	age	p_value	effect_size
Female	25 to 29	0.4991200	0.0007471	Male	25 to 29	0.6781376	0.0006325
Female	30 to 34	0.2494899	0.0011155	Male	30 to 34	0.2494899	0.0013071
Female	35 to 39	0.2494899	0.0008115	Male	35 to 39	0.2494899	0.0011484
Female	40 to 44	0.5329178	0.0002668	Male	40 to 44	0.2494899	0.0008268
Female	45 to 49	0.4046282	0.0003022	Male	45 to 49	0.2281511	0.0007286
Female	50 to 54	0.2080808	0.0004873	Male	50 to 54	0.2080808	0.0005453
Female	55 to 59	0.2080808	0.0004585	Male	55 to 59	0.1552752	0.0003804
Female	60 to 64	0.2080808	0.0003426	Male	60 to 64	0.1259214	0.0002925
Female	65 to 69	0.2080808	0.0001949	Male	65 to 69	0.1128847	0.0001938

- The alternative hypothesis (HA): There is a difference in the slope of the prevalence between the "before" period and the "after" period.

Because the sampling is random, we will use the two-tailed Wilcoxon rank sum test to test the hypotheses.

```
opioid_use_na2 = opioid_use_na %>%
  group_by(sex, age) %>%
  group_split() %>%
  map(~{
    before = diff(.$val[.$year < 1998])
    after = diff(.$val[.$year >= 1998])
    # calculate the slope
    tibble(
      sex = .$sex[1],
      age = .$age[1],
      p_value = wilcox.test(after, before)$p.value,
      effect_size = median(after) - median(before)
    )
  }) %>%
  bind_rows() %>%
  group_by(sex) %>%
  group_split()
kable(opioid_use_na2)
```

Table 3.1: The age (age), the sex (sex), the p value (p_value) and the effect size (effect_size) of the the Wilcoxon rank sum test for each group.

From Table 3.1, within each age and sex group, the p value of the Wilcoxon rank sum test is larger than 0.05, we cannot reject the null hypothesis (H0).

However, if we do not consider the sex as a covariate, we will take the average of the prevalence of male and female to represent the prevalence of the whole population within each age group, as the ratio of the male population and the female population within each age group of the divided locations is approximately 1:1.

```
opioid_use_na3 = opioid_use_na %>%
  group_by(year, age) %>%
  summarize(val = mean(val)) %>%
  group_by(age) %>%
  group_split() %>%
  map(~{
    before = diff(.$val[.$year < 1998])
    after = diff(.$val[.$year >= 1998])
```

```
    # calculate the slope
    tibble(
      age = .$age[1],
      p_value = wilcox.test(after, before)$p.value,
      effect_size = median(after) - median(before)
    )
  }) %>%
  bind_rows()
kable(opioid_use_na3)
```

age	p_value	effect_size
25 to 29	0.3212577	0.0001825
30 to 34	0.0000203	0.0004856
35 to 39	0.0000017	0.0003229
40 to 44	0.0038462	0.0001567
45 to 49	0.0000507	0.0001326
50 to 54	0.0000017	0.0001998
55 to 59	0.0000034	0.0002215
60 to 64	0.0000760	0.0001585
65 to 69	0.0001115	0.0001016

Table 3.2: The age (age), the p value (p_value) and the effect size (effect_size) of the the Wilcoxon rank sum test for each group.

From Table 3.2, when we do not take the sex as a covariate, we can identify that except 25-29 age group, the p-values for the other age groups are all smaller than 0.05, so we can reject the null hypothesis (H0). Therefore, except 25-29 age group, there is sufficient evidence to conclude that there is significant difference in the slope of the prevalence between the "before" period and the "after" period. Additionally, as the effect size (after - before) are all larger than 0, so we can confirm that the increasing prescription of the opioid can accelerate the increase in the prevalence of the opioid use disease except 25-29 age group.

From Figure 3.1 and Figure 3.2, we can assume that there is a linear relationship between the year and the prevalence. We use the linear regression model to fit the data, and we test the normality and the homoscedastic of the residuals. To find which age group is the most affected, we compare the difference in the slope of the "before" and "after" period except 25-29 age group.

As the "Multiple R-squared" represents the proportion of the variability of the dependent variable explained by the model, we set the threshold of the "Multiple R-squared" as 0.7 (recommended in the ADS2 Lecture 2.6 Slide Page 29).

```
opioid_use_na4 = opioid_use_na %>%
  filter(age != "25 to 29") %>%
  group_by(year, age) %>%
  summarize(val = mean(val)) %>%
  group_by(age) %>%
  group_split() %>%
  map(~{
    before_data = filter(., year < 1998)
    after_data = filter(., year >= 1998)
    before = summary(lm(val ~ year, before_data))
    after = summary(lm(val ~ year, after_data))
    tibble(age = .$age[1],
           before_r_squ = before$r.squared,
```

```
            before_slope = before$coefficients[2],
            after_r_squ = after$r.squared,
            after_slope = after$coefficients[2],
            slope_diff = after_slope - before_slope
            )
  }) %>%
  bind_rows() %>%
  arrange(desc(slope_diff))
kable(opioid_use_na4)
```

age	before_r_squ	before_slope	after_r_squ	after_slope	slope_diff
30 to 34	0.9989009	0.0005917	0.9623719	0.0012381	0.0006464
35 to 39	0.9989153	0.0003462	0.9462965	0.0008878	0.0005417
40 to 44	0.9974509	0.0003236	0.9371975	0.0006411	0.0003175
50 to 54	0.9940458	0.0001388	0.9680568	0.0004446	0.0003058
55 to 59	0.9740971	0.0000632	0.9715578	0.0003448	0.0002816
45 to 49	0.9964566	0.0002517	0.9460636	0.0005158	0.0002641
60 to 64	0.7917317	0.0000218	0.9640205	0.0002345	0.0002127
65 to 69	0.4042371	0.0000093	0.9481171	0.0001471	0.0001378

Table 3.3: The age (age), the "Multiple r-squared" of the "before" period (before_r_squ), the slope of the prevalence in the "before" period (before_slope), the "Multiple r-squared" of the "after" period (after_r_squ), the slope of the prevalence in the "after" period (after_slope), and the difference in the slope between the "before" and "after" period (slope_diff).

From Table 3.3, because the "Multiple r-squared" values are all greater than 0.7 except 65-69 age group (the "before" period), we can compare the difference in the slope between the "before" and "after" period except 65-69 age group. The 30-34 age group has the largest difference in the slope of the prevalence of the opioid use disease, so the 30-34 age group is the most affected.

You can use the following code to further verify the normality and homoscedastic of the residuals of this model. Due to the page limit, we do not show the results of the code.

```
plot(model, c(1, 2))
```

Conclusion

If we take the sex as a covariate, we cannot confirm the increasing prescription of the opioid can accelerate the increase in the prevalence of the opioid use disease in each age group. If we do not take the sex as a covariate, we can confirm that the increasing prescription of the opioid can accelerate the increase in the prevalence of the opioid use disease except 25-29 age group, and the 30-34 age group is the most affected.

Part 2: Ask your own question

First, in Figure 2.1, we identify that in east Asia and Pacific, for all age groups, the alcohol-related disease prevalence of the male is higher than that of the female during the years. We are interested in whether in different locations in the world, the alcohol-related death rate of the male is higher than the female. Then, we do some visualization.

```r
data1 = substance_use %>%
  filter(measure == "Deaths",
         cause == "Alcohol use disorders") %>%
  select(location, age, year, sex, val)

data1$location = gsub(data1$location, pattern = " - WB", replacement = "")

ggplot(data1, aes(x = location, y = val, fill = location)) +
  geom_boxplot(outlier.size = 0.1) +
  labs(x = "Age", y = "Difference in death rate (%)") +
   theme_bw() +
  facet_wrap(~sex) +
  theme(axis.text.x = element_blank(),
        strip.background =  element_rect(fill = NA, colour = NA))
```

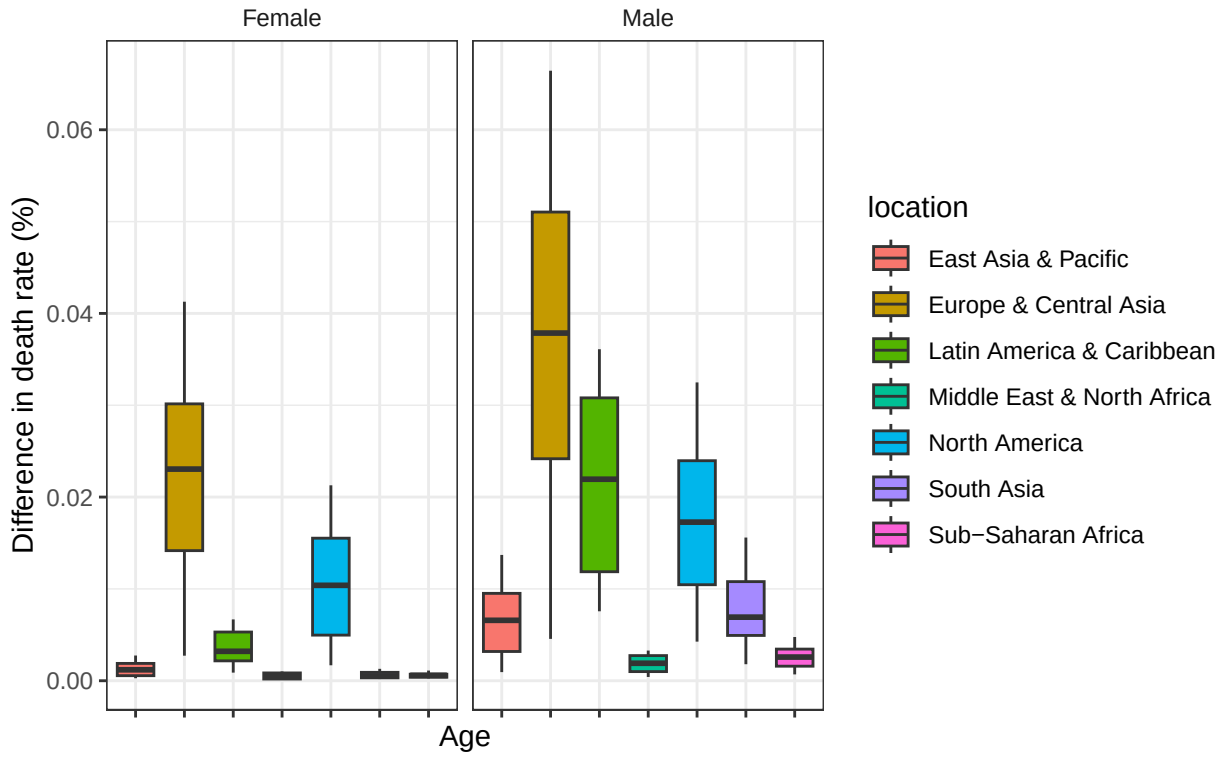


Figure 4.1 The the alcohol-related death rate of the male and the female in different locations in the world.

From Figure 4.1, we find that the alcohol-related death rates in Sub-Saharan Africa and Middle East & North Africa are much lower than the other locations for both male and female, so there is little significance to compare the difference in death rate in these two locations. we regard them as outliers and drop them out. Here we give a possible explanation: both Sub-Saharan Africa and Middle East & North Africa are mostly "underdeveloped areas". People in these places have few chances to have a heavy alcohol exposure.

Next, we compare the difference in the death rate of the male and female in the selected locations.

```r
locations = c("East Asia & Pacific", "North America", "South Asia",
"Latin America & Caribbean", "Europe & Central Asia")

data2 = data1 %>%
```

```
  filter(location %in% locations) %>%
  group_by(location, age, year) %>%
  spread(key = "sex", val = "val") %>%
  summarize(diff = Male - Female) %>%
  bind_rows()

ggplot(data2, aes(x = age, y = diff, fill = age)) +
  geom_boxplot(outlier.size = 0.1) +
  facet_wrap(~location, scale = "free_y") +
  labs(x = "Age", y = "Difference in death rate (%)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position = "none",
        strip.background =  element_rect(fill = NA, colour = NA))
```

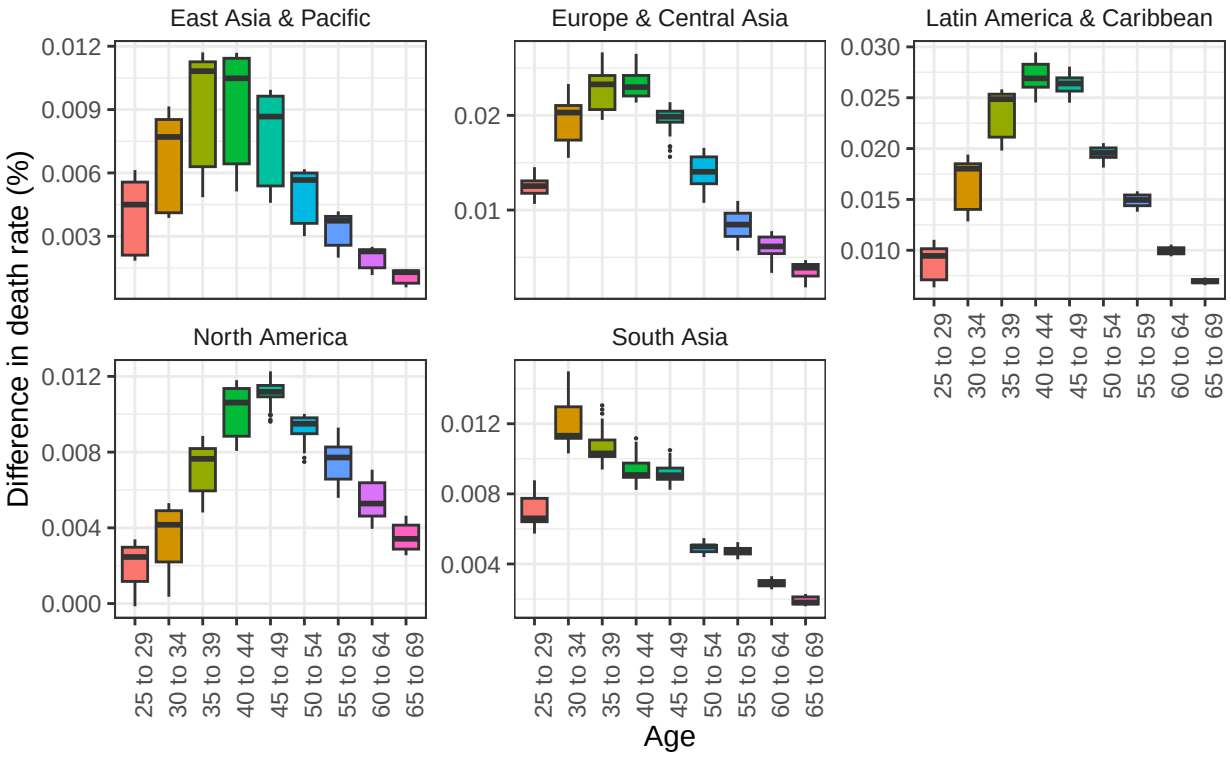


Figure 4.2 The difference in the alcohol-related death rate between male and female in different locations.

From Figure 4.2, we find that in different locations, the differences in the prevalence between the male and the female in different age groups have different distributions. We are interested in whether the location and the age have any interaction when they affect the difference in the death rate.

Therefore, our question is

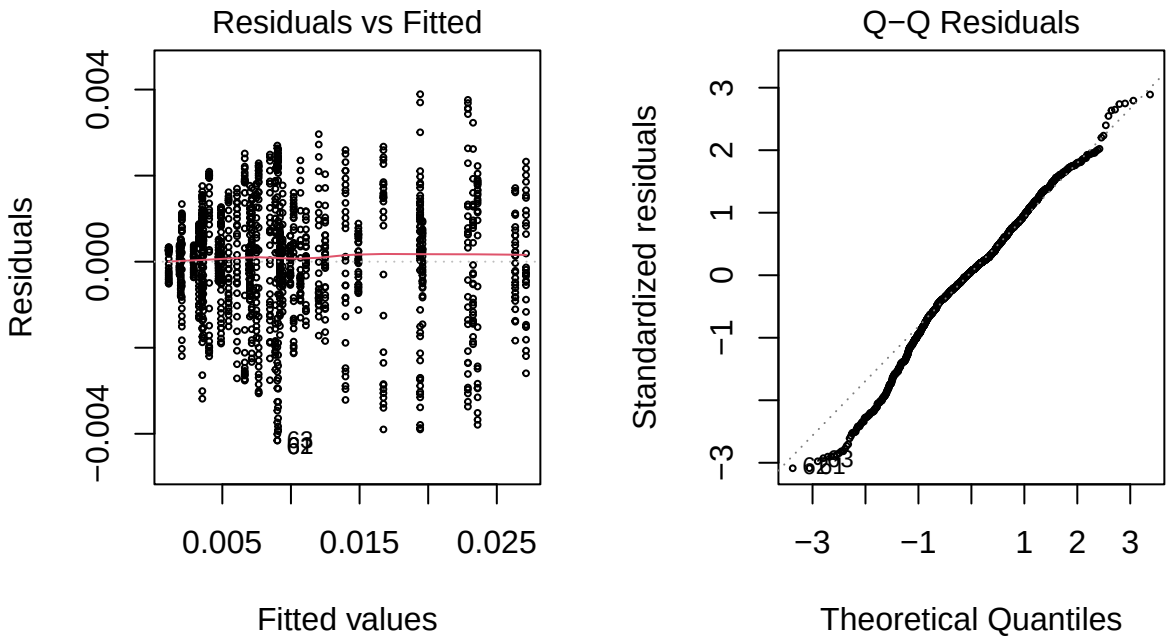
Do the location and the age have any interaction when they affect the difference in the alcohol-related death rate between the male and the female?

First, we formulate the hypotheses.

- The null hypothesis (H0): There is no interaction between the location and the age.
- The alternative hypothesis (HA): There is an interaction between the location and the age.

First, we build a model. As the sampling is random, we test the normality and homoscedastic of the residuals.

```
model2 = lm(diff ~ location * age, data = data2)
par(mfrow = c(1, 2))
plot(model2, c(1, 2),cex=0.4)
```




```
par(mfrow = c(1, 1))
summary(model2)$r.squared
```

```
## [1] 0.9635071
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: diff
##                Df    Sum Sq   Mean Sq F value    Pr(>F)
## location        4 0.0304297 0.0076074 4055.99 < 2.2e-16 ***
## age             8 0.0241903 0.0030238 1612.17 < 2.2e-16 ***
## location:age   32 0.0100047 0.0003126  166.69 < 2.2e-16 ***
## Residuals    1305 0.0024477 0.0000019
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Q-Q plot, we can find that the residuals are normally distributed. From the Residuals vs Fitted plot, we can find that the residuals are homoscedastic.

The "Multiple R-squared" value represents the proportion of the variability of the dependent variable explained by the model. The "Multiple R-squared" value is 0.9635071, so the model explains the data quite well. The p-value of ANOVA is smaller than 2.2e-16, so we can reject the null hypothesis (H0). There is sufficient evidence to conclude that there is a significant interaction between the location and the age.

Conclusion

There is a significant interaction between the location and the age when they affect the difference in the alcohol-related death rate between the male and the female.