

# ADS2 Group Exercise ICA

Group 6

2024-04-02

## 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

First, we need to extract the data of the alcohol-related death rate among men aged 40-44.

We use pipeline operations to construct the code, which avoids excess intermediate variables, enhances readability and makes the logical relationship clearer.

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

highest_alcohol_deaths = substance_use %>%
  filter(measure == "Deaths", year == 2019, age == "40 to 44",
         sex == "Male", cause == "Alcohol use disorders")
  # Filter the data

highest_alcohol_deaths
```

```
##   measure                location sex    age                cause
## 1 Deaths                South Asia - WB Male 40 to 44 Alcohol use disorders
## 2 Deaths Middle East & North Africa - WB Male 40 to 44 Alcohol use disorders
## 3 Deaths                East Asia & Pacific - WB Male 40 to 44 Alcohol use disorders
## 4 Deaths                North America Male 40 to 44 Alcohol use disorders
## 5 Deaths                Sub-Saharan Africa - WB Male 40 to 44 Alcohol use disorders
## 6 Deaths                Europe & Central Asia - WB Male 40 to 44 Alcohol use disorders
## 7 Deaths Latin America & Caribbean - WB Male 40 to 44 Alcohol use disorders
##   metric year      val      upper      lower
## 1 Percent 2019 0.012215856 0.014481335 0.008484016
## 2 Percent 2019 0.003040330 0.003688087 0.002506647
## 3 Percent 2019 0.012726958 0.014213882 0.008809356
## 4 Percent 2019 0.029002889 0.031514494 0.026391834
## 5 Percent 2019 0.003210615 0.004246450 0.002634772
## 6 Percent 2019 0.053798538 0.058466137 0.047957598
## 7 Percent 2019 0.032451149 0.034494076 0.030397480
```

In this dataset, there are not two rows with identical location but different measurement values. However, if another dataset has such rows, we decided to calculate the average death rate as the the measurement value for each location.

```
highest_alcohol_deaths = highest_alcohol_deaths %>%
  group_by(location) %>%
  # Group by "location" so that the average death rate can be calculated
  # separately for each location
  summarize(average_death_rate = mean(val)) %>%
  # The data for each location were aggregated and the average death rate was calculated
  arrange(desc(average_death_rate)) %>%
  # Sort by average death rate in descending order
  top_n(1, average_death_rate)
  # Select the region with the highest average death rate

highest_alcohol_deaths
```

```
## # A tibble: 1 x 2
##   location                average_death_rate
##   <chr>                  <dbl>
## 1 Europe & Central Asia - WB      0.0538
```

## 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 from different years, age groups and sex groups, and calculate the average prevalence rate in case of more than one values with identical attributes.

```
eap_alcohol_data = substance_use %>%
  filter(measure == "Prevalence", cause == "Alcohol use disorders",
         location == "East Asia & Pacific - WB")
eap_alcohol_trends = eap_alcohol_data %>%
  group_by(year, age, sex) %>%
  summarize(average_prevalence = mean(val, na.rm = TRUE)) %>%
  # The change of prevalence over time and age group was analyzed and compared by sex
  ungroup()
  # Remove the group status
```

Next, we visualize the data from different years, age groups and sex groups. We visualize the data in two ways, faceting the plots by age groups and by sex groups. We do not mix the data together by using the average number, because populations from different age groups and sex groups are different.

```

#eap_alcohol_trends1 = eap_alcohol_trends %>% paste0(.$age, " years old")
eap_alcohol_trends1 = eap_alcohol_trends
eap_alcohol_trends1$age = paste0(eap_alcohol_trends1$age, " years old")
ggplot(eap_alcohol_trends1,
      aes(x = year, y = average_prevalence, color = sex,
          group = interaction(sex, age))) +
  geom_line() +
  facet_wrap(~age, scales = 'free_y') +
  # Use the panel diagram to show the different age groups
  labs(x = "Year",
       y = "Average Prevalence (%)",
       color = "Sex") +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5)
  )

```

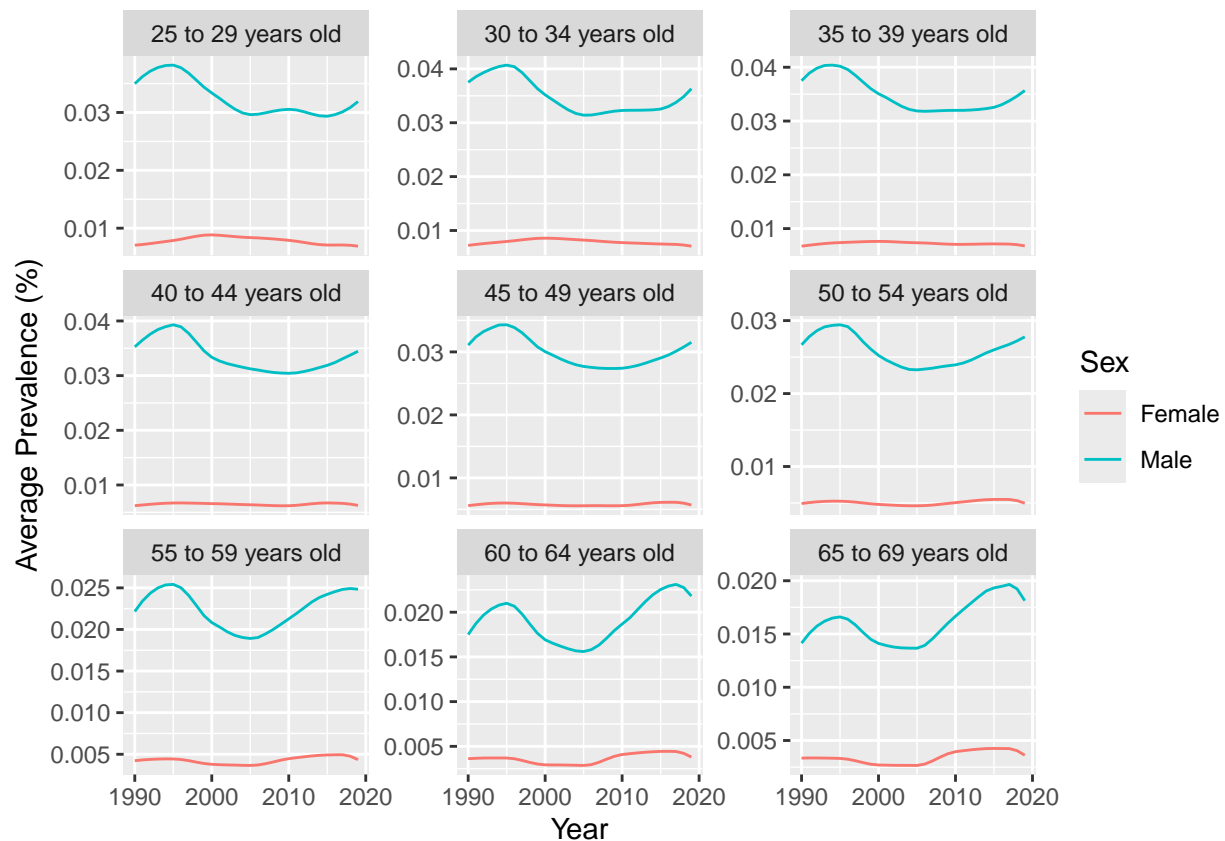


Figure 1: The trends of alcohol-related disease prevalence in east Asia and Pacific. The plots in this figure are faceted by age groups, better for visually comparing the prevalence between the male and the female.

```

ggplot(eap_alcohol_trends,
      aes(x = year, y = average_prevalence, color = age,
          group = interaction(sex, age))) +
  geom_line() +
  facet_wrap(~sex, scales = 'free_y') +
  # Use the panel diagram to show the different age groups
  labs(x = "Year",

```

```

y = "Average Prevalence (%)",
color = "Age (years old)" +
theme(legend.position = "bottom",
      legend.title.position = "top",
      legend.title = element_text(hjust = 0.5),
      plot.title = element_text(hjust = 0.5)
)

```

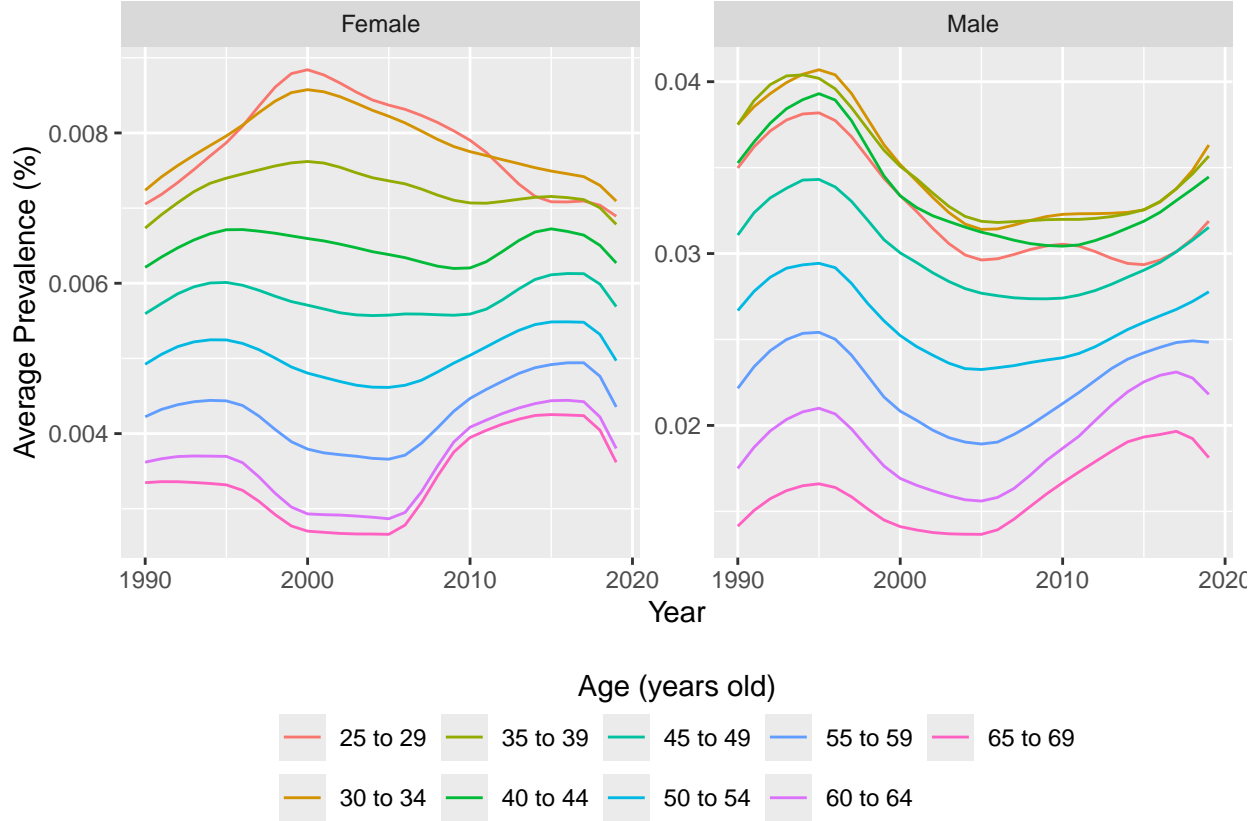


Figure 2: The trends of alcohol-related disease prevalence in east Asia and Pacific. The plots in this figure are faceted by sex groups, better for visually comparing the prevalence between different age groups.

From the two plots, you can see how the prevalence of alcohol-related disease in the East Asia and Pacific region and in different age groups. Also, you can clearly identify different patterns of the prevalence between male and female.

Further, to validate the difference between male and female more statistically. First, we formulate the hypotheses.

- The null hypothesis ( $H_0$ ): There is no difference in the prevalence of different age groups between male and female.
- The alternative hypothesis ( $H_A$ ): There is difference in the prevalence of different age groups between male and female.

The data for male and female in the same year are correlated and paired (i.e. observational data under the same conditions), and we want to test the differences in prevalence between male and female in each year, so we will use the paired statistical test in the following. We use the Shapiro test to test the normality of the differences in prevalence between male and female.

- 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 test to test the hypotheses.

```
eap_alcohol_data2 = eap_alcohol_data %>%
  select(year, age, sex, val) %>%
  group_by(age) %>%
  group_split() %>%
  map(~spread(., key = "sex", value = "val")) %>%
  # Convert to the wide format
  map(~{
    shapiro_p = shapiro.test(.$Male - .$Female)$p.value
    if(shapiro_p < 0.05){
      test_type = "Wilcoxon test"
      p_value = wilcox.test(.$Male, .$Female, paired = T)$p.value
    }
    else{
      test_type = "Student's t-test"
      p_value = t.test(.$Male, .$Female, paired = T)$p.value
    }
    tibble(shapiro_p = shapiro_p,
            test_type = test_type,
            p_value = p_value)
  }) %>%
  # perform statistical test in all age groups
  bind_rows() %>%
  mutate(age = unique(eap_alcohol_trends$age), .before = shapiro_p)
eap_alcohol_data2
```

```
## # A tibble: 9 x 4
##   age      shapiro_p test_type      p_value
##   <chr>      <dbl> <chr>      <dbl>
## 1 25 to 29  0.000199 Wilcoxon test 0.00000000186
## 2 30 to 34  0.00128  Wilcoxon test 0.00000000186
## 3 35 to 39  0.000252 Wilcoxon test 0.00000000186
## 4 40 to 44  0.000635 Wilcoxon test 0.00000000186
## 5 45 to 49  0.00239  Wilcoxon test 0.00000000186
## 6 50 to 54  0.00440  Wilcoxon test 0.00000000186
## 7 55 to 59  0.0289   Wilcoxon test 0.00000000186
## 8 60 to 64  0.0411   Wilcoxon test 0.00000000186
## 9 65 to 69  0.0185   Wilcoxon test 0.00000000186
```

## Conclusion

Of all the age groups, the p values from the statistical test are all less than  $10^{-8}$ . Therefore, we reject the null hypothesis( $H_0$ ). There is sufficient evidence to support the conclusion that there are significant differences 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

Since there is no data named “United States” in the location column, we extract the data from “North America”. Because the problem requires us to analyze the data since the late 1990s, we extract the data since 1998. As the population of the male and the female are not known to be equal, the data cannot be mixed by simply calculating the average number.

```
opioid_use_na = substance_use %>%
  filter(measure == "Prevalence",
         location == "North America",
         cause == 'Opioid use disorders',
         year >= 1998)
opioid_trends_na = opioid_use_na %>%
  group_by(year, age, sex) %>%
  summarize(average_prevalence = mean(val)) %>%
  ungroup()
```

Next, we visualize the data from different years, age groups and sex groups. We visualize the data in two ways, faceting the plots by age groups and by sex groups.

```
opioid_trends_na1 = opioid_trends_na
opioid_trends_na1$age = paste0(opioid_trends_na1$age, " years old")
ggplot(opioid_trends_na1,
       aes(x = year, y = average_prevalence, color = sex,
           group = interaction(sex, age))) +
  geom_line() +
  facet_wrap(~age, scales = 'free_y') +
  # Use the panel diagram to show the different age groups
  labs(x = "Year",
       y = "Average Prevalence (%)",
       color = "Sex") +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45)
  )
```

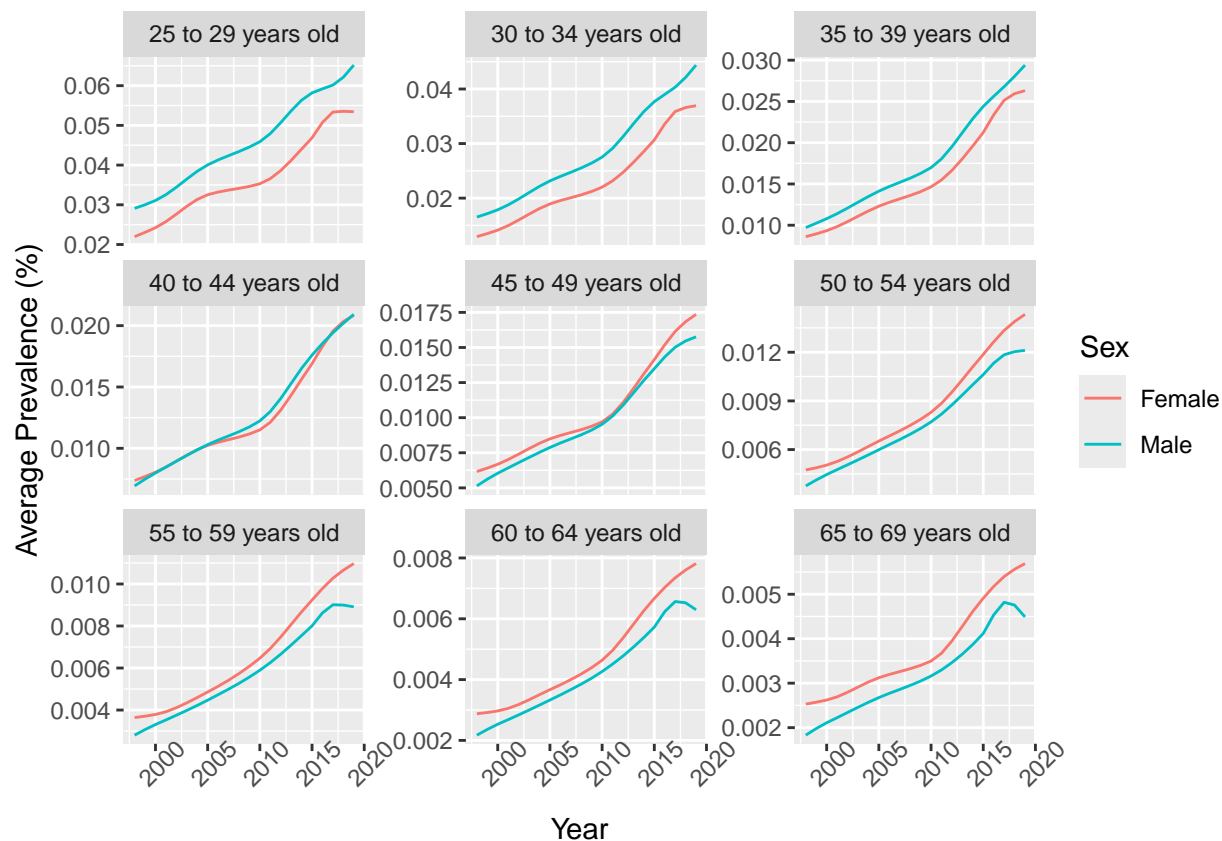


Figure 3: The trends of opioid use related disease prevalence in North America. The plots in this figure are faceted by age groups, better for visually comparing the prevalence between the male and the female.

```
ggplot(opioid_trends_na,
  aes(x = year, y = average_prevalence, color = age,
    group = interaction(sex, age))) +
  geom_line() +
  facet_wrap(~sex, scales = 'free_y') +
  # Use the panel diagram to show the different age groups
  labs(x = "Year",
    y = "Average Prevalence (%)",
    color = "Age (years old)") +
  theme(legend.position = "bottom",
    legend.title.position = "top",
    legend.title = element_text(hjust = 0.5),
    plot.title = element_text(hjust = 0.5))
```

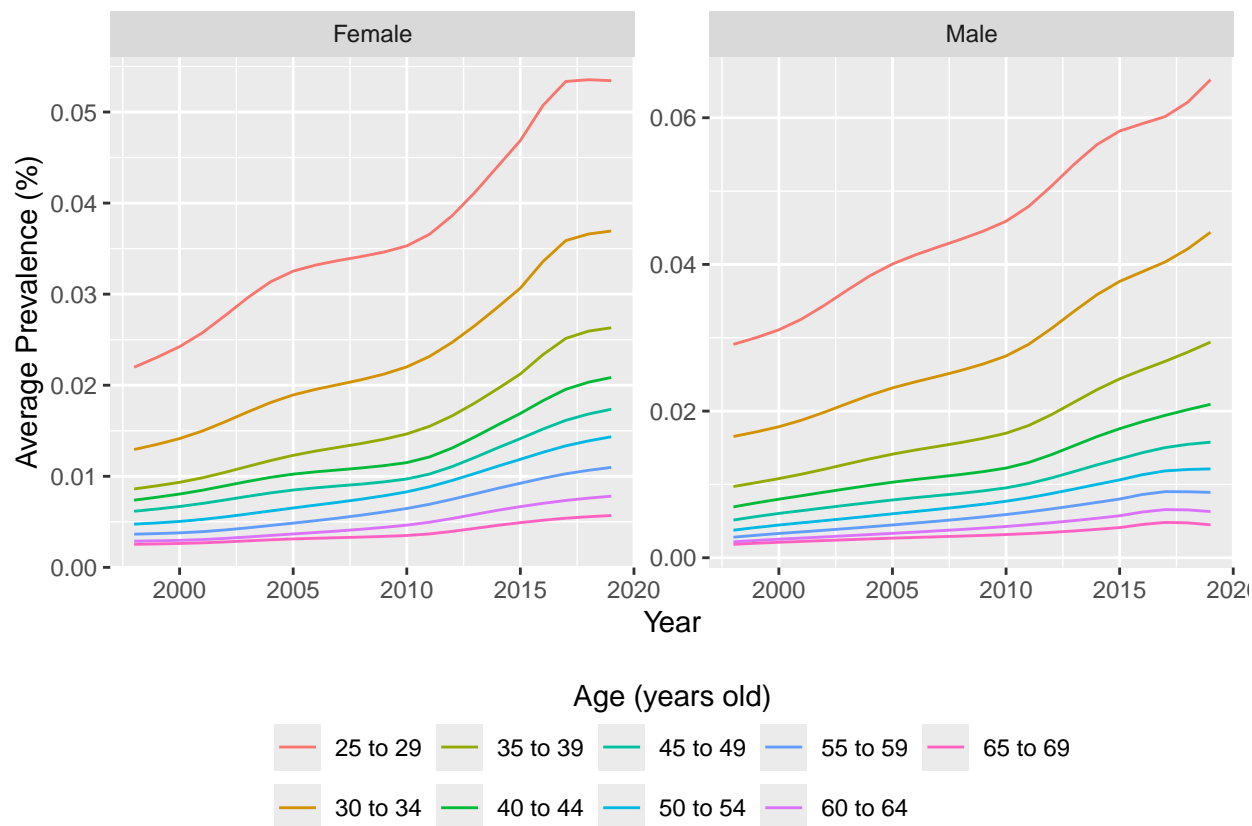


Figure 4: The trends of opioid use related disease prevalence in North America. The plots in this figure are faceted by sex groups, better for visually comparing the prevalence between different age groups.

```

opioid_trends_na2 = opioid_use_na %>%
  select(year, age, sex, val) %>%
  group_by(age, sex) %>%
  group_split() %>%
  map(~{
    model = lm(val ~ year, .)
    shapiro_p = shapiro.test(model$residuals)$p.value
    test_type = if(shapiro_p < 0.05) "spearman" else "pearson"
    res = cor.test(.$year, .$val, method = test_type)
    tibble(age = .$age[1],
           sex = .$sex[1],
           shapiro_p = shapiro_p,
           test_type = test_type,
           p_value = res$p.value,
           cor = res$estimate)
  }) %>%
  # perform statistical test in all age groups
  bind_rows() %>%
  group_by(sex) %>%
  group_split() %>%
  map(~{arrange(., desc(cor))})
#kable(opioid_trends_na2)
opioid_trends_na2

```



```
## [[1]]
## # A tibble: 9 x 6
##   age      sex shapiro_p test_type p_value cor
##   <chr>   <chr>    <dbl> <chr>    <dbl> <dbl>
## 1 55 to 59 Female    0.136 pearson  2.92e-15 0.979
## 2 25 to 29 Female    0.240 pearson  5.87e-15 0.977
## 3 50 to 54 Female    0.172 pearson  7.16e-15 0.977
## 4 60 to 64 Female    0.218 pearson  2.80e-14 0.973
## 5 30 to 34 Female    0.418 pearson  2.85e-14 0.973
## 6 35 to 39 Female    0.237 pearson  4.09e-13 0.965
## 7 65 to 69 Female    0.190 pearson  9.95e-13 0.962
## 8 45 to 49 Female    0.232 pearson  1.08e-12 0.962
## 9 40 to 44 Female    0.245 pearson  3.80e-12 0.956
##
## [[2]]
## # A tibble: 9 x 6
##   age      sex shapiro_p test_type p_value cor
##   <chr>   <chr>    <dbl> <chr>    <dbl> <dbl>
## 1 25 to 29 Male      0.138 pearson  1.59e-21 0.995
## 2 55 to 59 Male      0.124 pearson  8.15e-19 0.991
## 3 50 to 54 Male      0.241 pearson  2.21e-18 0.990
## 4 60 to 64 Male      0.0811 pearson  1.94e-17 0.987
## 5 30 to 34 Male      0.385 pearson  8.35e-17 0.985
## 6 45 to 49 Male      0.0848 pearson  6.32e-16 0.982
## 7 65 to 69 Male      0.166 pearson  1.62e-15 0.980
## 8 35 to 39 Male      0.327 pearson  3.90e-15 0.978
## 9 40 to 44 Male      0.0812 pearson  5.74e-15 0.977
```

Table 3: The age group (age), the sex group (sex), the p value of the Shapiro test (shapiro\_p), the correlation test type (test\_type), the p value of the correlation test, and the correlation coefficient (cor).

From this table, you can identify that the p values for the correlation tests are all less than  $10^{-11}$ , and the correlation coefficients are all greater than 0.95, so we can confirm an increase in the prevalence of diseases related to opioid use in North America since 1998.

You can also use

```
plot(model, 2)
```

to further verify whether the residuals of this model are homoscedastic.

From this table, you can also identify that

- For the female, the 55 to 59 age group has the highest correlation coefficient, so this group is the most affected group.
- For the male, the 25 to 29 age group has the highest correlation coefficient, so this group is the most affected group.