AssignmentI

Group2

2024-11-19

Introduction

Data Exploration

Question 1

Overview :

Creating a plot showing the number of cases by age group and sex.

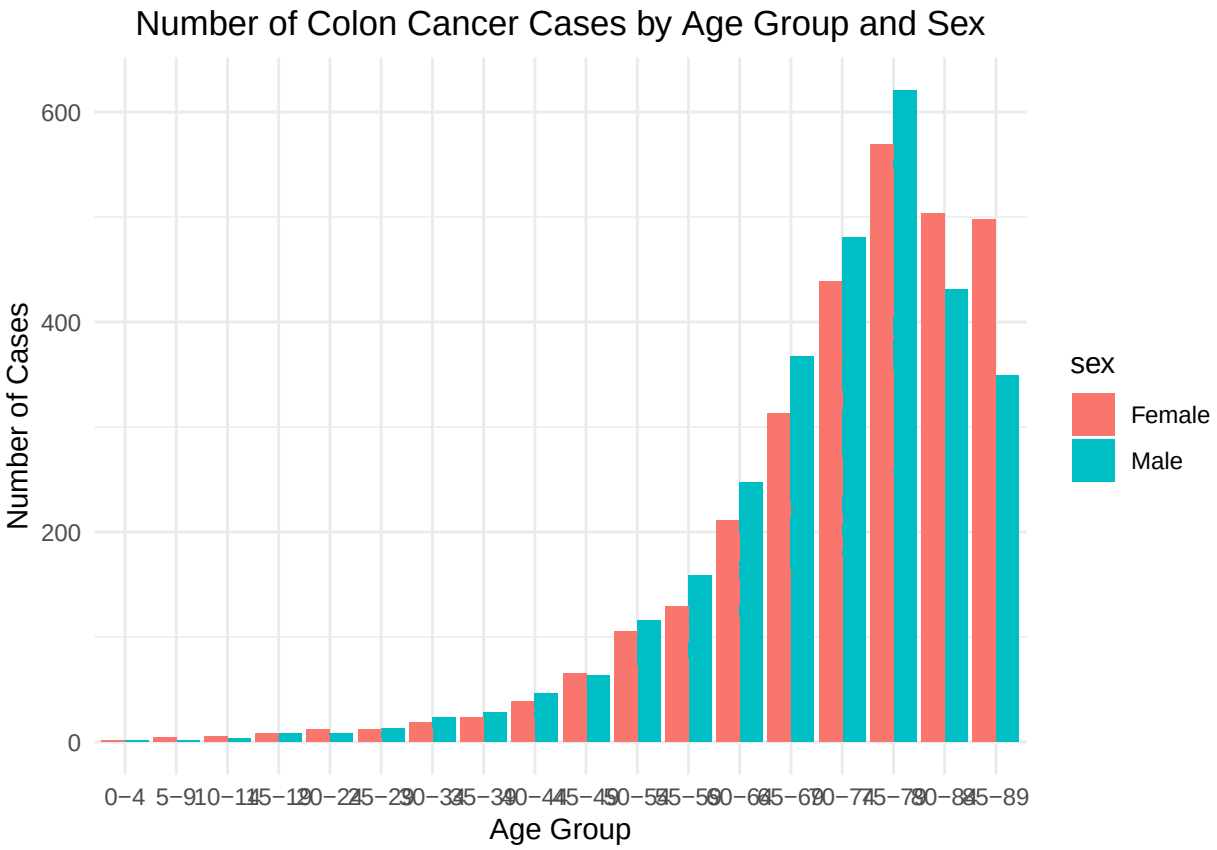
Code :

```
path_to_cases <- "cases.tsv"
cases <- read_tsv(path_to_cases)
```

```
## Rows: 1908 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): agegroup, sex
## dbl (2): year, n
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
cases$agegroup <- factor(cases$agegroup,
            levels = c("0-4", "5-9", "10-14", "15-19", "20-24",
            "25-29", "30-34", "35-39", "40-44", "45-49",
            "50-54", "55-59", "60-64", "65-69", "70-74",
            "75-79", "80-84", "85-89"))

ggplot(cases, aes(x = agegroup, y = n, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Colon Cancer Cases by Age Group and Sex",
       x = "Age Group",
       y = "Number of Cases") +
  theme_minimal()+
  scale_y_continuous(labels = label_comma())+
  theme(plot.title = element_text(hjust = 0.5))
```



Output :

Generated plot visualizing age group distribution.

Analysis :

From the plot we can conclude that the risk of getting a colon cancer increases exponentially above the age approx. 40-44, for both men and women. We can also see a clear trend of the number of cases being slightly higher for men up to the age group 80-84 when the trend reverses.

Question 2

Overview :

Plotting colon cancer cases across subsequent years.

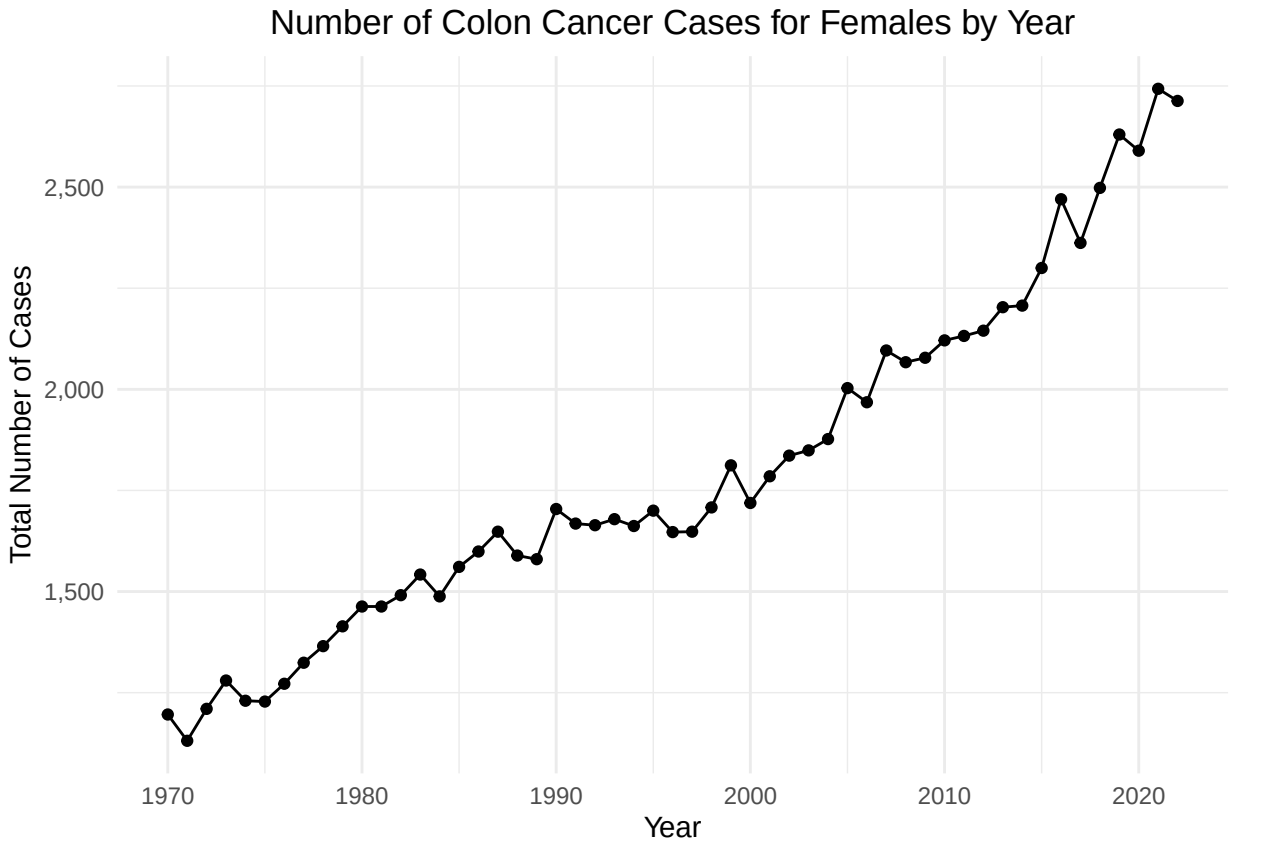
Code :

```
total_cases_by_year_sex<- cases %>%
  group_by(year, sex) %>%
  summarise(total_cases = sum(n, na.rm = TRUE))
```

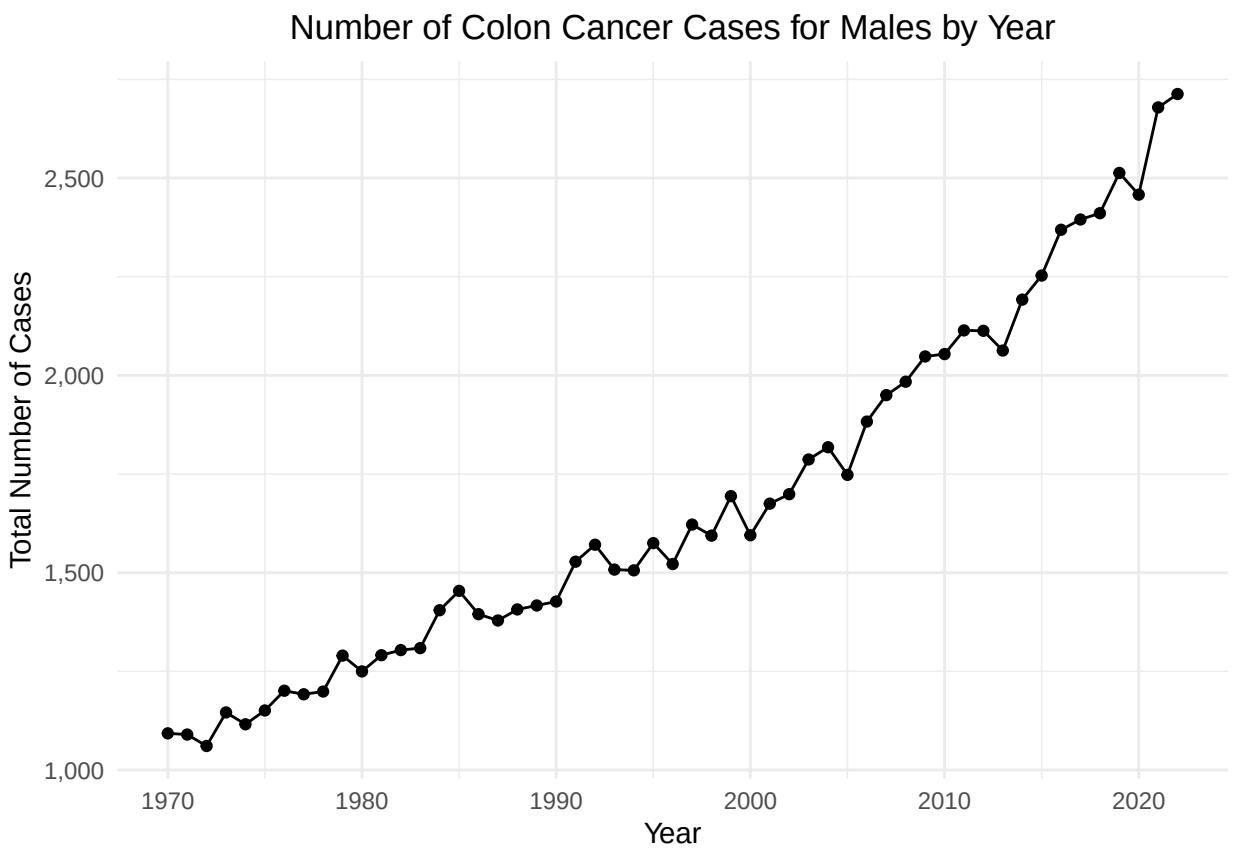
```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
ggplot(subset(total_cases_by_year_sex, sex == "Female"), aes(x = year, y = total_cases)) +
  geom_line() +
```

```
  geom_point() +
  labs(title = "Number of Colon Cancer Cases for Females by Year",
       x = "Year",
       y = "Total Number of Cases") +
  theme_minimal()+
  scale_y_continuous(labels = label_comma())+
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(subset(total_cases_by_year_sex, sex == "Male"), aes(x = year, y = total_cases)) +
  geom_line() +
  geom_point() +
  labs(title = "Number of Colon Cancer Cases for Males by Year",
       x = "Year",
       y = "Total Number of Cases") +
  theme_minimal()+
  scale_y_continuous(labels = label_comma())+
  theme(plot.title = element_text(hjust = 0.5))
```



Output :

Two plots showing the nunber of cases in each calendar year.

Analysis :

From the two graphs we can see that there exists a trend of increasing colon cancer cases in subsequent calendar years.

Question 3

Overview :

Exploring population data and plotting population size by age group and sex across subsequent years.

Code :

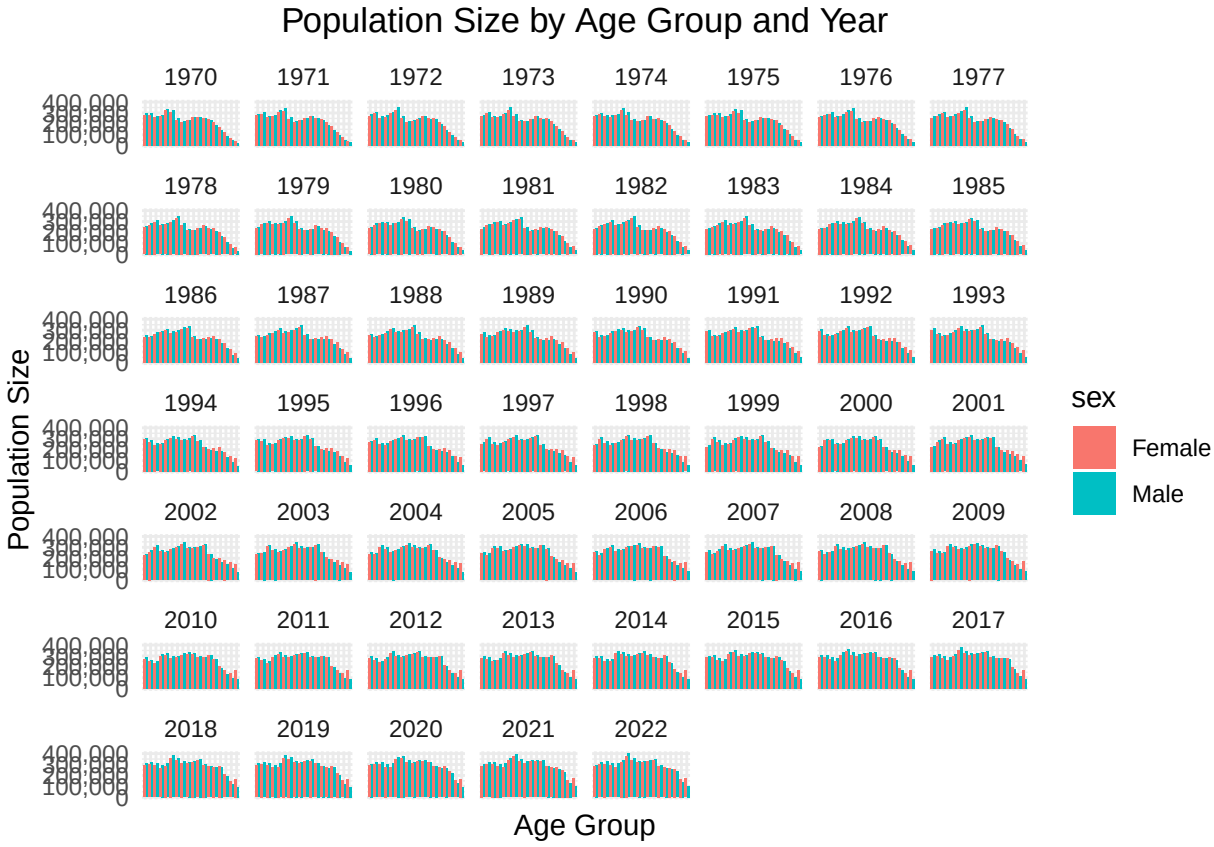
```
path_to_population <- "population.tsv"
population<- read_tsv(path_to_population)
```

```
## Rows: 1908 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): agegroup, sex
## dbl (2): year, n_pop
##
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
population$agegroup <- factor(population$agegroup,
                              levels = c("0-4", "5-9", "10-14", "15-19", "20-24",
                                         "25-29", "30-34", "35-39", "40-44", "45-49",
                                         "50-54", "55-59", "60-64", "65-69", "70-74",
                                         "75-79", "80-84", "85-89"))


ggplot(population, aes(x = agegroup, y = n_pop, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ year) +
  labs(title = "Population Size by Age Group and Year",
       x = "Age Group",
       y = "Population Size") +
  theme_minimal() +
  scale_y_continuous(labels = label_comma())+
  theme(axis.text.x = element_blank())+
  theme(plot.title = element_text(hjust = 0.5))
```



Output :

A figure showing a plots for every year in the data set. Each of the plots shows pupulation size across different age groups, the bars colors correspond to sex (orange - female, green - male).

Analysis :

Population file inculdes the same variables and their types as cases file. The age groups, and calendar years are also the same. The difference is in the order of columns, and the year column is reversed in population, relative to cases.

Question 4

Overview :

Merging the data and creating a data frame that shows total number of cases and the total population of males and females in each year.

Code :

```
merged_data <- left_join(cases, population, by = c("agegroup", "year", "sex"))

summary_data <- merged_data %>%
  group_by(year, sex) %>%
  summarise(
    total_cases = sum(n, na.rm = TRUE),
    total_population = sum(n_pop, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
head(summary_data)
```

```
## # A tibble: 6 x 4
## # Groups:   year [3]
##    year sex    total_cases total_population
##   <dbl> <chr>        <dbl>            <dbl>
## 1  1970 Female        1196          4045318
## 2  1970 Male          1093          4035911
## 3  1971 Female        1131          4066592
## 4  1971 Male          1090          4048573
## 5  1972 Female        1210          4077814
## 6  1972 Male          1061          4051315
```

Output :

New data frame that stores information merged from both of the analyzed data sets.

Question 5

Overview :

Creating a new data frame that will show incidence rate among both sexes in each year.

Code :

```
merged_data <- merged_data %>%
  mutate(incidence_rate = n/ n_pop)
```

```r
head(merged_data)
```

```
## # A tibble: 6 x 6
##   agegroup  year sex       n  n_pop incidence_rate
##   <fct>    <dbl> <chr> <dbl>  <dbl>          <dbl>
## 1 0-4       2022 Male      0 296183     0
## 2 5-9       2022 Male      0 319820     0
## 3 10-14     2022 Male      1 325003     0.00000308
## 4 15-19     2022 Male      8 310539     0.0000258
## 5 20-24     2022 Male      5 310354     0.0000161
## 6 25-29     2022 Male      4 342974     0.0000117
```

```r
summary_data <- summary_data %>%
  mutate(incidence_rate = total_cases / total_population)

head(summary_data)
```

```
## # A tibble: 6 x 5
## # Groups:   year [3]
##    year sex    total_cases total_population incidence_rate
##   <dbl> <chr>        <dbl>            <dbl>          <dbl>
## 1  1970 Female        1196          4045318       0.000296
## 2  1970 Male          1093          4035911       0.000271
## 3  1971 Female        1131          4066592       0.000278
## 4  1971 Male          1090          4048573       0.000269
## 5  1972 Female        1210          4077814       0.000297
## 6  1972 Male          1061          4051315       0.000262
```

Output :

A dataframe with information on the number of cases, populaltion and incidence year for men and women in each year.

Analysis :

Incidence rate is the number of new cases of the outcome divided by the total person-time at risk, for a specific follow-up period. Here we simply divide the number of cases by the population size, without accounting for the time. For this type of data where we do not know the person-time at risk, it appears to be an appropriate way of calculating an incidence rate. However, this way should provide a reasonable estimate when, for example, the population is relatively stable over the time period (we can see from the plots in Question 3 that the population changes over the given time - we have more people in the elderly age group as the years increase). This suggests that this data is not suitable to infere the incidence rate.

Question 6

Overview :

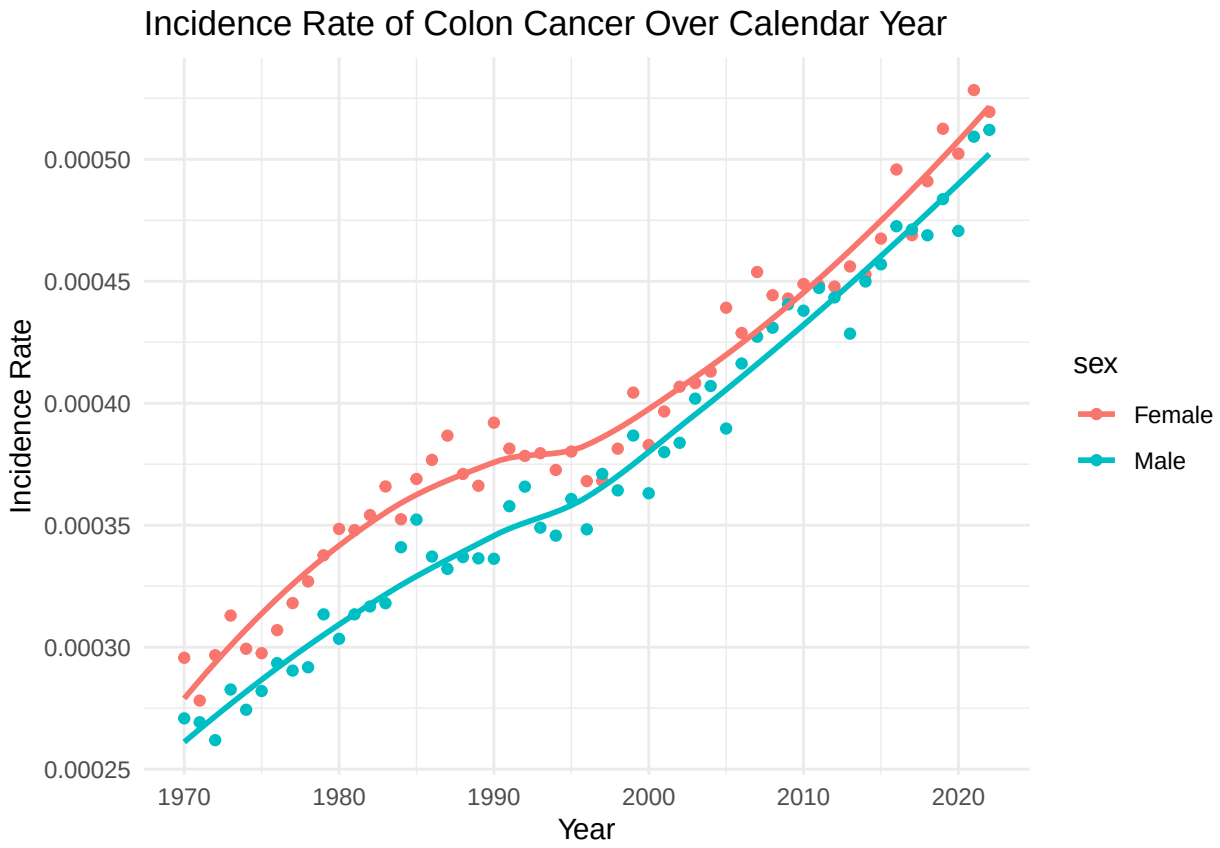
Plotting the incidence rate of colon cancer over calendar time and apply asmoother, separately by males and females. Creating a graph of incidence rates over calendar year by sex and age group, and apply smoothers.

Code :

```r
ggplot(summary_data, aes(x = year, y = incidence_rate, color = sex)) +
  geom_point() +
```

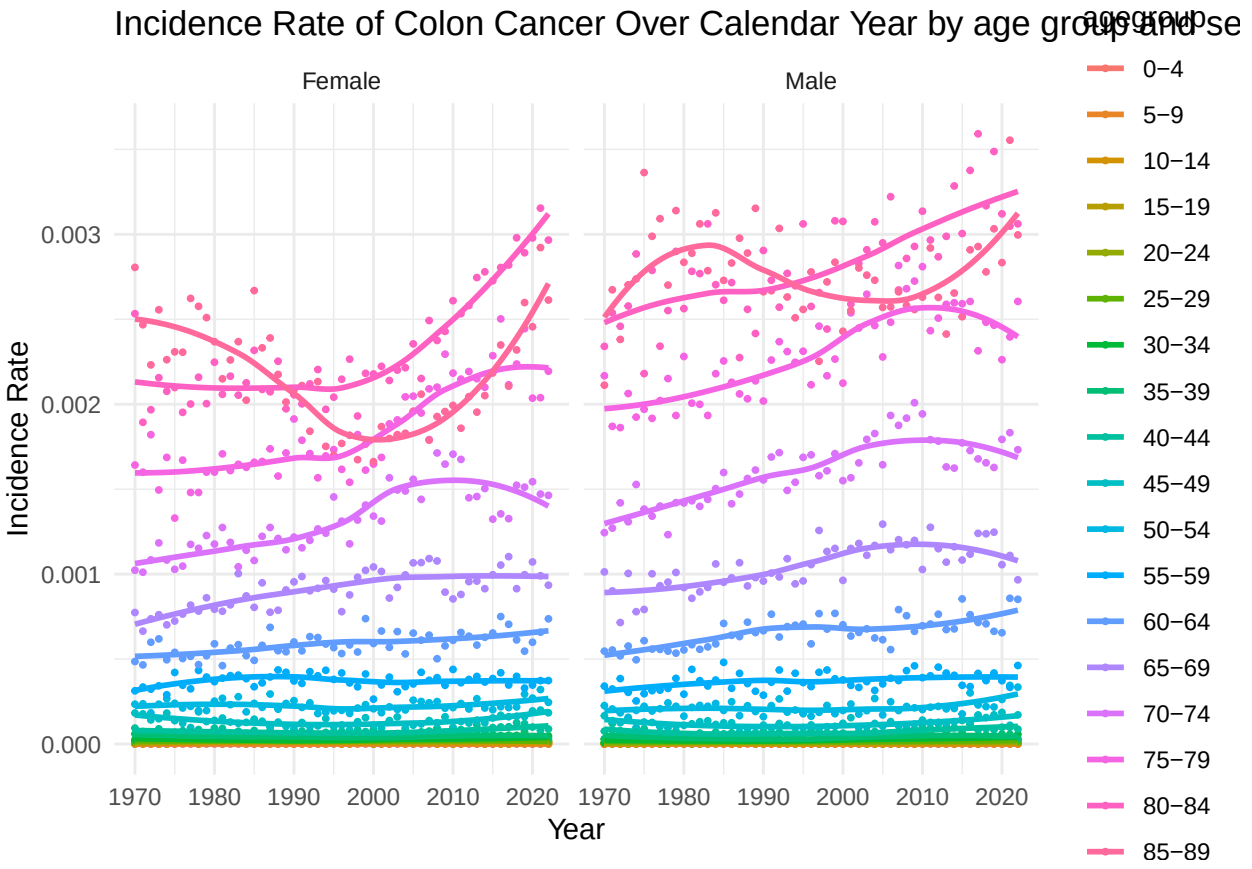
```
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Incidence Rate of Colon Cancer Over Calendar Year",
       x = "Year",
       y = "Incidence Rate") +
  theme_minimal()
```

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




```
ggplot(merged_data, aes(x = year, y = incidence_rate, color = agegroup)) +
  geom_point(cex = 0.7) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Incidence Rate of Colon Cancer Over Calendar Year by age group and sex",
       x = "Year",
       y = "Incidence Rate") +
  theme_minimal() +
  facet_wrap(~ sex)
```

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



Output :

A graph showing the smoothed incidence rate trends for males and females over time. Faceted graphs showing incidence rate trends by age group and sex over calendar years.

Analysis :

Overall Trends: The smoothed curves reveal increasing incidence rates for both sexes over the observed time frame. Differences in incidence rates between males and females are consistent over time, with females generally having higher rates.

Age Group-Specific Trends:There is a noticeable rise in incidence rates over time in most age groups, particularly in the older populations. The youngest age groups (e.g., 0-4, 5-9, 10-14) maintain consistently low incidence rates throughout the years, with little variation. For males, the incidence rates in the highest age groups (e.g., 70-74 and above) are more pronounced compared to females. The higher the age group, the higher the probability of colon cancer, but for age group 85-89, the incidence of colon cancer fluctuated with the increase of years, and after around 1990, the incidence of colon cancer in this age group was lower than that in age group 80-84 The trends for middle-aged groups (e.g., 40-44, 50-54) exhibit moderate increases over the years, with smoother and less steep curves compared to older groups. This suggests a gradual risk increase with age but highlights that the exponential increase occurs predominantly in the elderly.

Question 7

Overview :

Fitting a suitable Poisson model with the total number of cases as dependent variable, using the population size as an offset, and calendar year and sex as independent variables.

Code :

```
poisson_model <- glm(total_cases ~ year + sex + offset(log(total_population)),
                     data = summary_data,
                     family = poisson)

summary(poisson_model)
```

```
##
## Call:
## glm(formula = total_cases ~ year + sex + offset(log(total_population)),
##     family = poisson, data = summary_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.969e+01  3.071e-01  -96.68   <2e-16 ***
## year         1.094e-02  1.536e-04   71.25   <2e-16 ***
## sexMale     -5.592e-02  4.658e-03  -12.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5505.07  on 105  degrees of freedom
## Residual deviance:  222.81  on 103  degrees of freedom
## AIC: 1211.4
##
## Number of Fisher Scoring iterations: 3
```

Output :

A detailed summary of the Poisson regression model, showing the coefficients, standard errors, z-values, and significance levels for each predictor.

Analysis :

Yearly Trend: The positive coefficient for year suggests a consistent increase in colon cancer incidence rates over time. Sex-Specific Differences: While males generally have higher crude incidence rates than females, the negative coefficient for sexMale suggests that, after controlling for year and population size, the adjusted incidence rate is slightly lower for males.

Question 8

Overview :

Based on the poisson regression model, calculate the incidence rates in 1970 and 2020 for males and females, and discuss the assumptions about how rates change over time and between sexes.

Code :

```
predict_data <- data.frame(year = c(1970, 2020),
                           sex = rep(c("Male", "Female"), each = 2),
                           total_population = 100000)
```

```
predict_data$predicted_cases <- predict(poisson_model,
                                        newdata = predict_data,
                                        type = "response")

predict_data$incidence_rate <- (predict_data$predicted_cases / predict_data$total_population)
print(predict_data)
```

```
##   year    sex total_population predicted_cases incidence_rate
## 1 1970   Male            1e+05        27.73751   0.0002773751
## 2 2020   Male            1e+05        47.94199   0.0004794199
## 3 1970 Female            1e+05        29.33271   0.0002933271
## 4 2020 Female            1e+05        50.69918   0.0005069918
```

Output :

A data frame with year, sex, and total population in the first three columns and the predicted number of cases as well as the corresponding incidence rate in the last two columns.

Analysis :

From the result, the incidence rate in 1970 among males and females was 0.0002773751 and 0.0002933271, separately. Moreover, the incidence rate in 2020 among males and females was 0.0004794199 and 0.0004794199, separately. The assumption for year is that the incidence rate changes linearly over the calendar year, and the assumption for the sex is that there is no interaction between sex and year, which means that the ratio of the incidence rate of male and female is an unchanged constant.

Question 9

Overview :

Fit a Poisson regression model that adjusts for age groups and calculate the incidence rates for males and females in the 70-74 age group for 1970 and 2020.

Code :

```
poisson_model_age <- glm(n ~ year + sex + agegroup + offset(log(n_pop)),
                         data = merged_data,
                         family = poisson)

summary(poisson_model_age)
```

```
##
## Call:
## glm(formula = n ~ year + sex + agegroup + offset(log(n_pop)),
##     family = poisson, data = merged_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -27.338166   0.771867 -35.418  < 2e-16 ***
## year            0.005400   0.000155  34.851  < 2e-16 ***
## sexMale         0.155476   0.004687  33.172  < 2e-16 ***
## agegroup5-9     2.841194   0.727020   3.908 9.31e-05 ***
## agegroup10-14   4.304576   0.711734   6.048 1.47e-09 ***
## agegroup15-19   5.058536   0.709256   7.132 9.88e-13 ***
```

```
## agegroup20-24   5.263833   0.708782   7.427 1.11e-13 ***
## agegroup25-29   5.452471   0.708415   7.697 1.40e-14 ***
## agegroup30-34   5.912646   0.707926   8.352  < 2e-16 ***
## agegroup35-39   6.397141   0.707621   9.040  < 2e-16 ***
## agegroup40-44   6.962570   0.707402   9.842  < 2e-16 ***
## agegroup45-49   7.506562   0.707277  10.613  < 2e-16 ***
## agegroup50-54   8.052154   0.707205  11.386  < 2e-16 ***
## agegroup55-59   8.568783   0.707164  12.117  < 2e-16 ***
## agegroup60-64   9.092283   0.707139  12.858  < 2e-16 ***
## agegroup65-69   9.541911   0.707126  13.494  < 2e-16 ***
## agegroup70-74   9.944553   0.707120  14.063  < 2e-16 ***
## agegroup75-79  10.287697   0.707117  14.549  < 2e-16 ***
## agegroup80-84  10.504311   0.707121  14.855  < 2e-16 ***
## agegroup85-89  10.407244   0.707130  14.718  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 406971  on 1907  degrees of freedom
## Residual deviance:   3472  on 1888  degrees of freedom
## AIC: 12620
##
## Number of Fisher Scoring iterations: 6
```

```r
predict_data_age <- data.frame(year = c(1970, 2020),
                               agegroup = '70-74',
                               sex = rep(c("Male", "Female"), each = 2),
                               n_pop = 100000)

predict_data_age$predicted_cases <- predict(poisson_model_age,
                                            newdata = predict_data_age,
                                            type = "response")

predict_data_age$incidence_rate <- (predict_data_age$predicted_cases / predict_data_age$n_pop)
print(predict_data_age)
```

```
##   year agegroup    sex n_pop predicted_cases incidence_rate
## 1 1970    70-74   Male 1e+05        136.0602    0.001360602
## 2 2020    70-74   Male 1e+05        178.2354    0.001782354
## 3 1970    70-74 Female 1e+05        116.4685    0.001164685
## 4 2020    70-74 Female 1e+05        152.5708    0.001525708
```

Output :

A detailed summary information of the new poisson model. A data frame with year, age group, sex, and population in the first four columns and the predicted number of cases as well as the corresponding incidence rate in the last two columns.

Analysis :

From the result, the incidence rate in 1970 in age group 70-74 among males and females was 0.001360602 and 0.001164685, separately. Moreover, the incidence rate in 2020 in age group 70-74 among males and females was 0.001782354 and 0.001525708, separately.

Question 10

Overview :

Code :

Output :

Analysis :

Question 11

Overview :

The task is to calculate direct age-standardized incidence rates for colon cancer across calendar years and sexes, using the population age distribution from 2022 as the reference (standard population). The objective is to create a graph of the age-standardized incidence rates and compare it with the non-age-standardized (crude) rates for males and females over time.

Code :

To solve this, there are a series of steps that involves significant data pre-processing.

1. Define a Standard Population: Use the age distribution from 2022 as the reference.

```
# Summarize the age distribution in 2022
standard_population <- population %>%
  filter(year == 2022) %>%
  group_by(sex, agegroup) %>%
  summarise(Population_2022 = sum(n_pop), .groups = "drop")

head(standard_population)
```

```
## # A tibble: 6 x 3
##   sex    agegroup Population_2022
##   <chr>  <fct>              <dbl>
## 1 Female 0-4               280184
## 2 Female 5-9               301335
## 3 Female 10-14             306568
## 4 Female 15-19             292308
## 5 Female 20-24             275136
## 6 Female 25-29             323678
```

2. Calculate Crude Rates: Determine the number of cases per unit population for each year, age group, and sex.

```
# Merge cases and population data
merged_data <- cases %>%
  left_join(population, by = c("year", "agegroup", "sex"))

# Calculate crude incidence rates
merged_data <- merged_data %>%
  mutate(Crude_Incidence_Rate = n / n_pop)

head(merged_data)
```

```
## # A tibble: 6 x 6
##   agegroup  year sex       n  n_pop Crude_Incidence_Rate
```

```
##   <fct>    <dbl> <chr> <dbl>  <dbl>                <dbl>
## 1 0-4       2022 Male      0 296183           0
## 2 5-9       2022 Male      0 319820           0
## 3 10-14     2022 Male      1 325003           0.00000308
## 4 15-19     2022 Male      8 310539           0.0000258
## 5 20-24     2022 Male      5 310354           0.0000161
## 6 25-29     2022 Male      4 342974           0.0000117
```

3. Standardize Rates: Weight the crude rates by the 2022 age distribution to compute age-standardized rates.

```
standardized_rates <- merged_data %>%
  filter(!is.na(Crude_Incidence_Rate)) %>%
  left_join(standard_population, by = c("sex", "agegroup")) %>%
  group_by(year, sex) %>%
  summarise(
    Standardized_Rate = sum(Crude_Incidence_Rate * Population_2022) / sum(Population_2022),
    .groups = "drop"
  )

head(standardized_rates)
```

```
## # A tibble: 6 x 3
##    year sex    Standardized_Rate
##   <dbl> <chr>              <dbl>
## 1  1970 Female          0.000435
## 2  1970 Male            0.000398
## 3  1971 Female          0.000395
## 4  1971 Male            0.000397
## 5  1972 Female          0.000408
## 6  1972 Male            0.000383
```

4. Visualization: Compare age-standardized and crude incidence rates across calendar years to evaluate the impact of standardization and reveal true trends.

We first compute the crude incidence rates and group by year and sex, similar to how we computed 'standardized_rates'.

```
crude_rates <- merged_data %>%
  group_by(year, sex) %>%
  summarise(
    Crude_Rate = sum(n) / sum(n_pop),
    .groups = "drop"
  )
```

In order to compare the age-standardized rates vs the non-standardized rates, we combine the tables 'standardized_rates' and 'crude_rates' and create distinct columns to distinguish between each other.

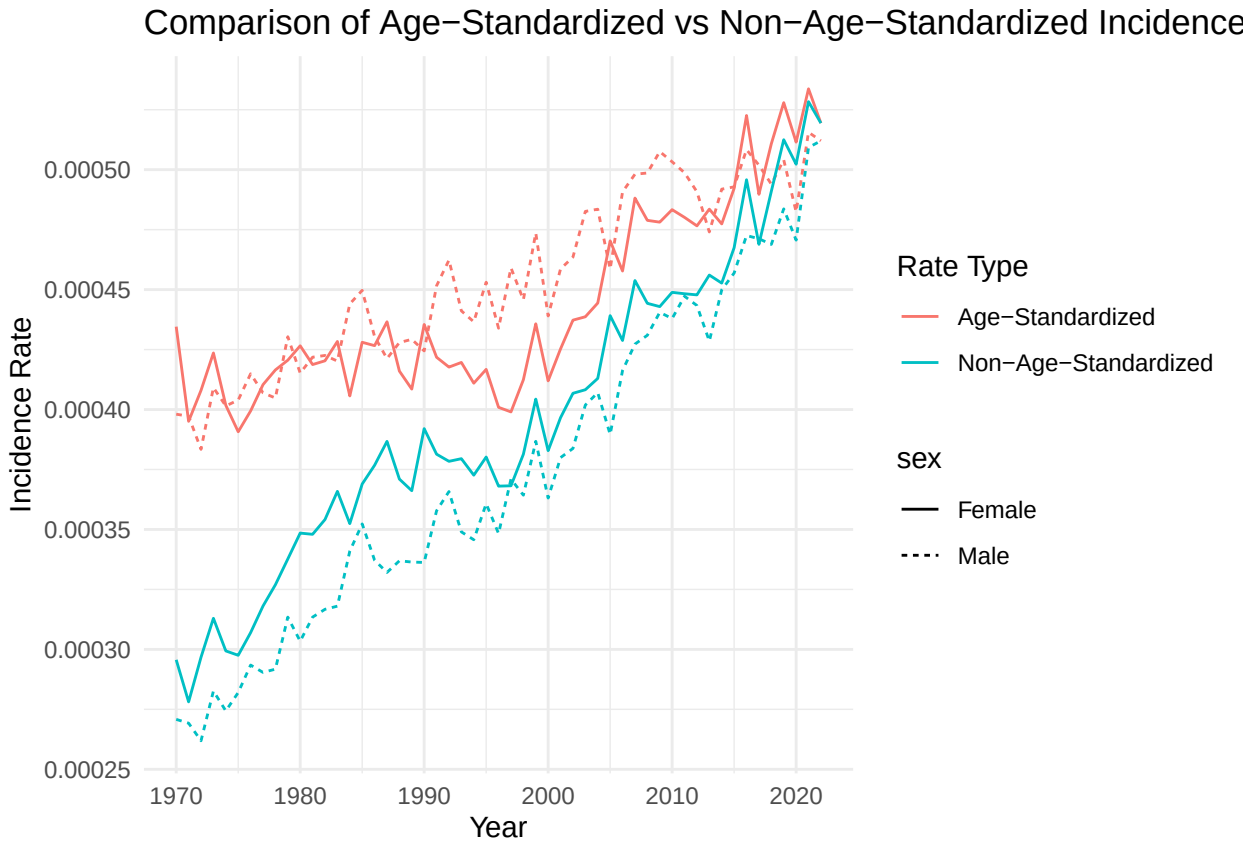
```
comparison_data <- crude_rates %>%
  rename(Rate = Crude_Rate) %>%
  mutate(Type = "Non-Age-Standardized") %>%
  bind_rows(
    standardized_rates %>%
      rename(Rate = Standardized_Rate) %>%
      mutate(Type = "Age-Standardized")
  )
```

```r
head(comparison_data)
```

```
## # A tibble: 6 x 4
##    year sex        Rate Type
##   <dbl> <chr>     <dbl> <chr>
## 1  1970 Female 0.000296 Non-Age-Standardized
## 2  1970 Male   0.000271 Non-Age-Standardized
## 3  1971 Female 0.000278 Non-Age-Standardized
## 4  1971 Male   0.000269 Non-Age-Standardized
## 5  1972 Female 0.000297 Non-Age-Standardized
## 6  1972 Male   0.000262 Non-Age-Standardized
```

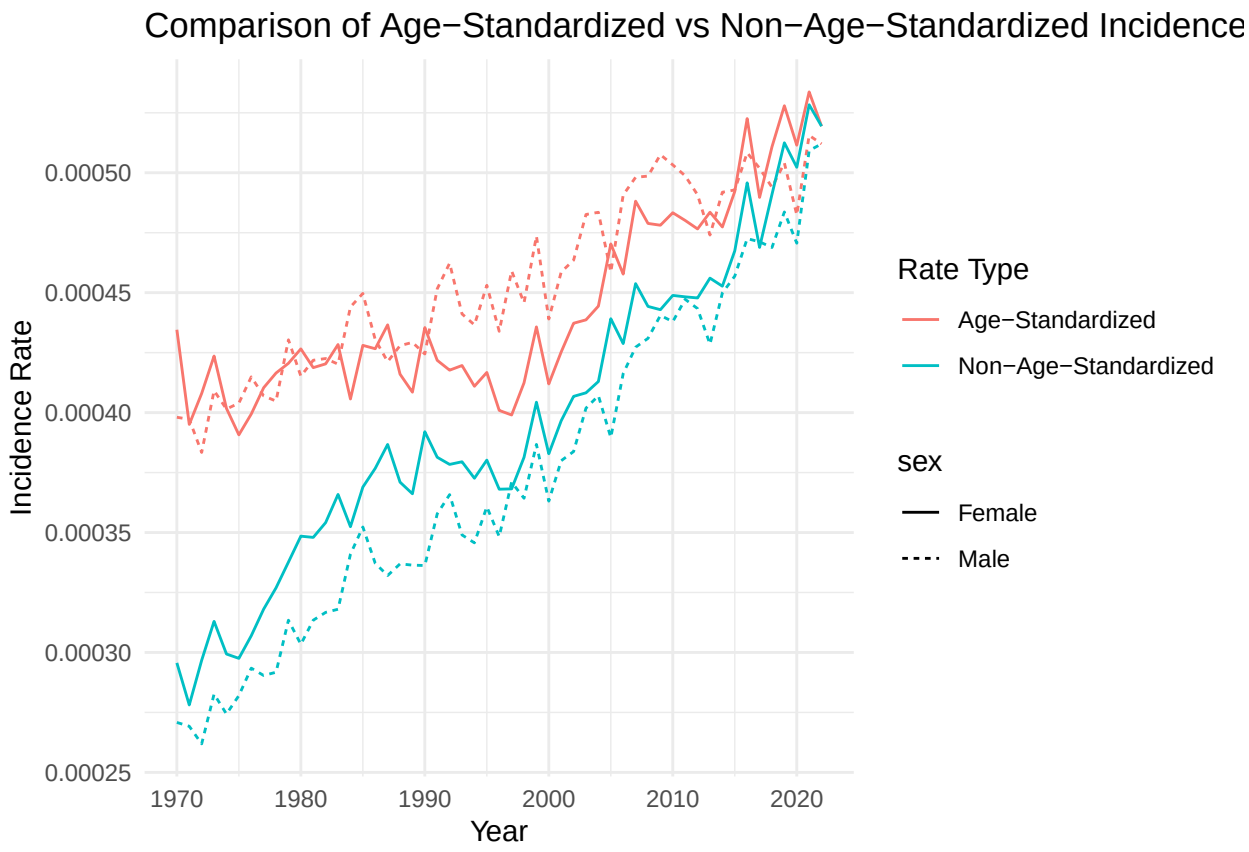
Finally, we plot the graph using this newly created table.

```r
ggplot(comparison_data, aes(x = year, y = Rate, color = Type, linetype = sex)) +
  geom_line() +
  labs(
    title = "Comparison of Age-Standardized vs Non-Age-Standardized Incidence Rates",
    x = "Year",
    y = "Incidence Rate",
    color = "Rate Type"
  ) +
  theme_minimal()
```



Output :

```r
ggplot(comparison_data, aes(x = year, y = Rate, color = Type, linetype = sex)) +
  geom_line() +
  labs(
    title = "Comparison of Age-Standardized vs Non-Age-Standardized Incidence Rates",
    x = "Year",
    y = "Incidence Rate",
    color = "Rate Type"
  ) +
  theme_minimal()
```



Analysis :

1. Difference Between Standardized and Crude Rates:

- The age-standardized rates are generally higher than the crude rates for both sexes, particularly in recent years.
- This suggests that the population's age structure has shifted, with an increasing proportion of older individuals (who have a higher incidence of colon cancer).

2. Trend Over Time:

- Both standardized and crude rates show an increasing trend over time, indicating a rising incidence of colon cancer from 1970 to 2020.
- The steeper increase in crude rates reflects the influence of population aging.

3. Sex-Specific Differences:

- Males consistently exhibit slightly higher incidence rates than females across all years in both age-standardized and crude measures.

- This could be attributed to differences in biological, behavioral, or exposure-related factors.

Conclusion:

Age-standardization adjusts for changes in the population's age distribution, providing a clearer picture of temporal trends. The increasing standardized rates suggest a true rise in colon cancer incidence beyond the effects of aging.

Question 12

Overview :

Code :

Output :

Analysis :

Question 13

Overview :

Code :

Output :

Analysis :