ki

Chen Chen Michelle

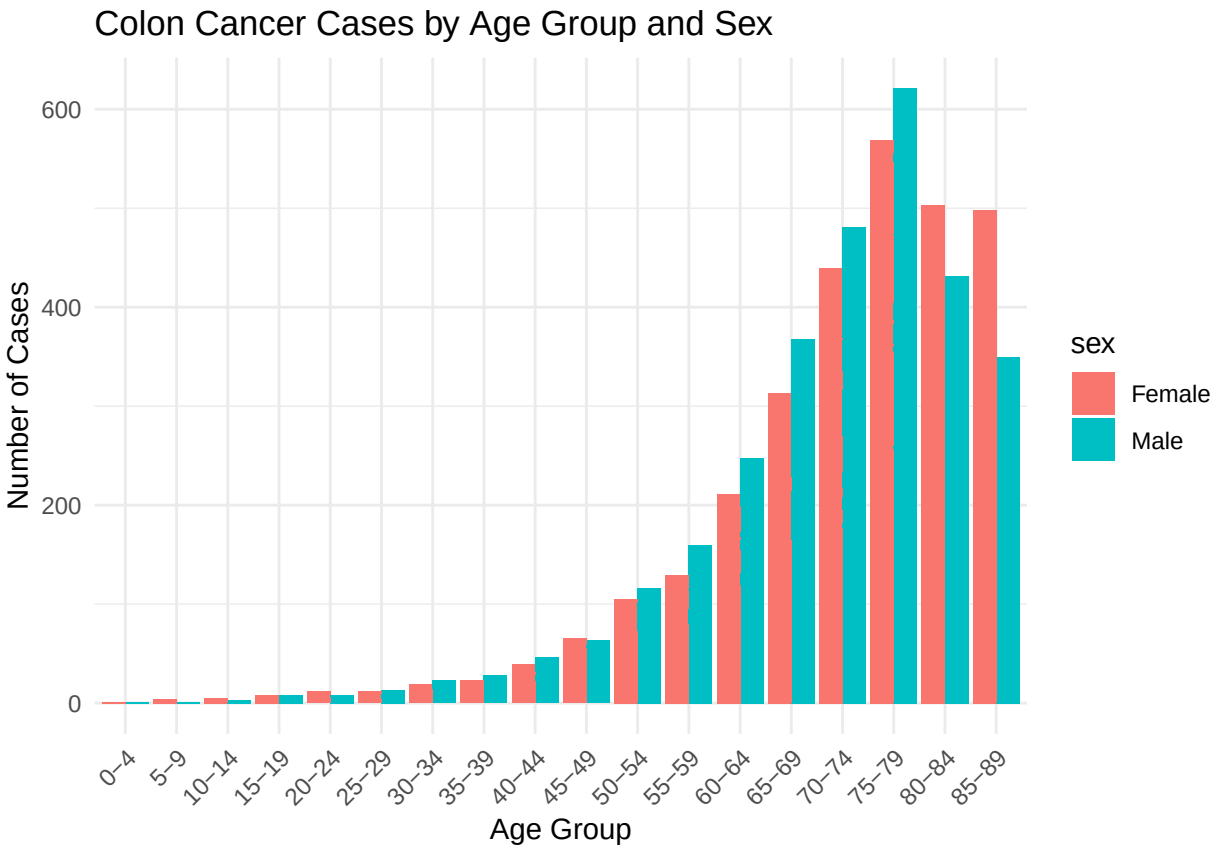
2024-11-18

We have variables are below: AgeGroup Year Sex The number of cases

Conclusion : when age increases, the number of the cancer cases are increasing too. The colon cancer tends to happen to the old aged group. We can guess age may be positively related to the colon cancer rate. In the plot, the age group 75-79 has the highest number of cases. And the age group over 80, the incidence rate tends to decrease.

```
# question 1
library(ggplot2)
data <- read.table("Downloads/cases.tsv", sep = "\t", header = TRUE)
data$agegroup <- factor(data$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(data, aes(x = agegroup, y = n, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Colon Cancer Cases by Age Group and Sex",
       x = "Age Group",
       y = "Number of Cases") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The trend of the total cases of both sexes has continuously increased over the years.

```
#question 2
library(dplyr)
```

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

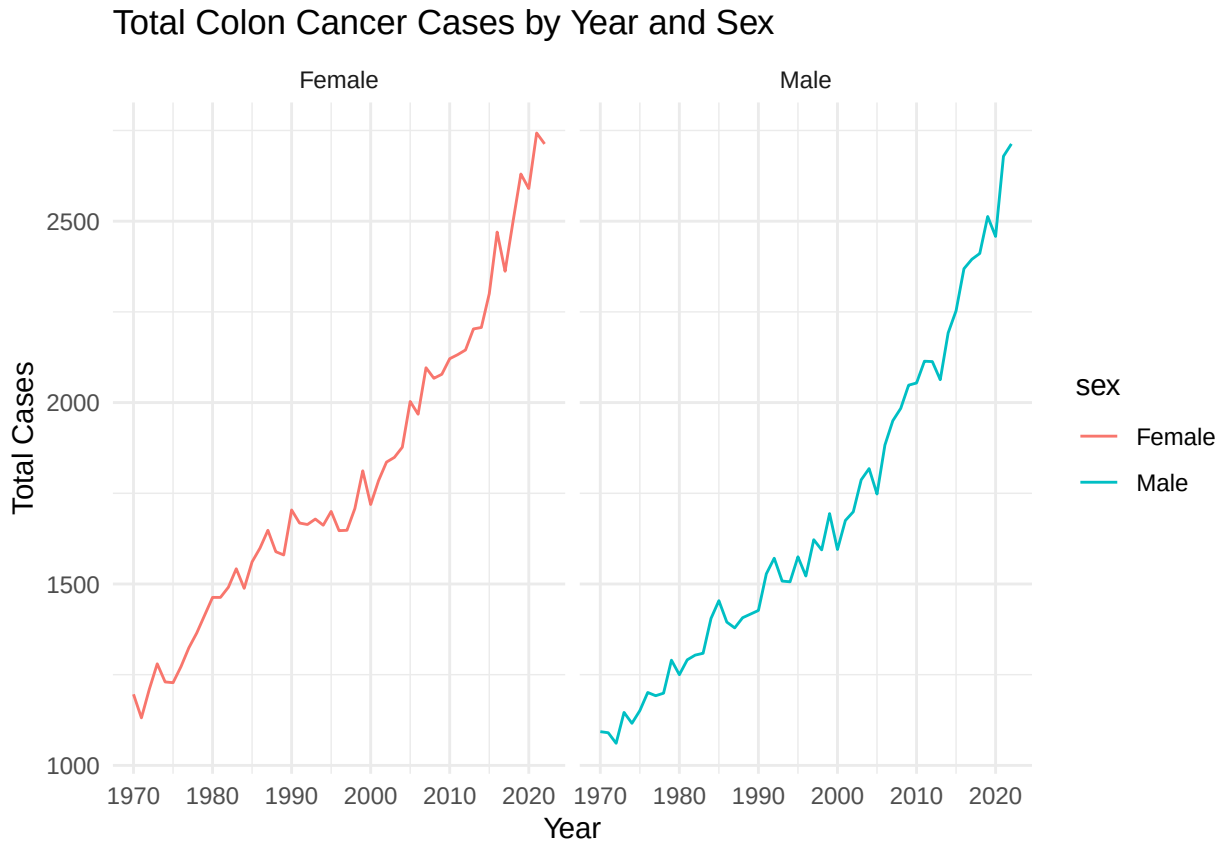
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
# Summarize the data
summary_data <- data %>%
  group_by(year, sex) %>%
  summarize(total_cases = sum(n, na.rm = TRUE))
```

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

```
ggplot(summary_data, aes(x = year, y = total_cases, colour  = sex)) +
  geom_line() +
  labs(title = "Total Colon Cancer Cases by Year and Sex",
       x = "Year",
       y = "Total Cases") +
  theme_minimal() +
  facet_wrap(~ sex)
```



The two datasets have the same age group and calendar years.

```
#question 3
population <- read.table("Downloads/population.tsv", sep = "\t", header = TRUE)
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"))
# Compare age groups
all(unique(data$agegroup) %in% unique(population$agegroup))
```

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

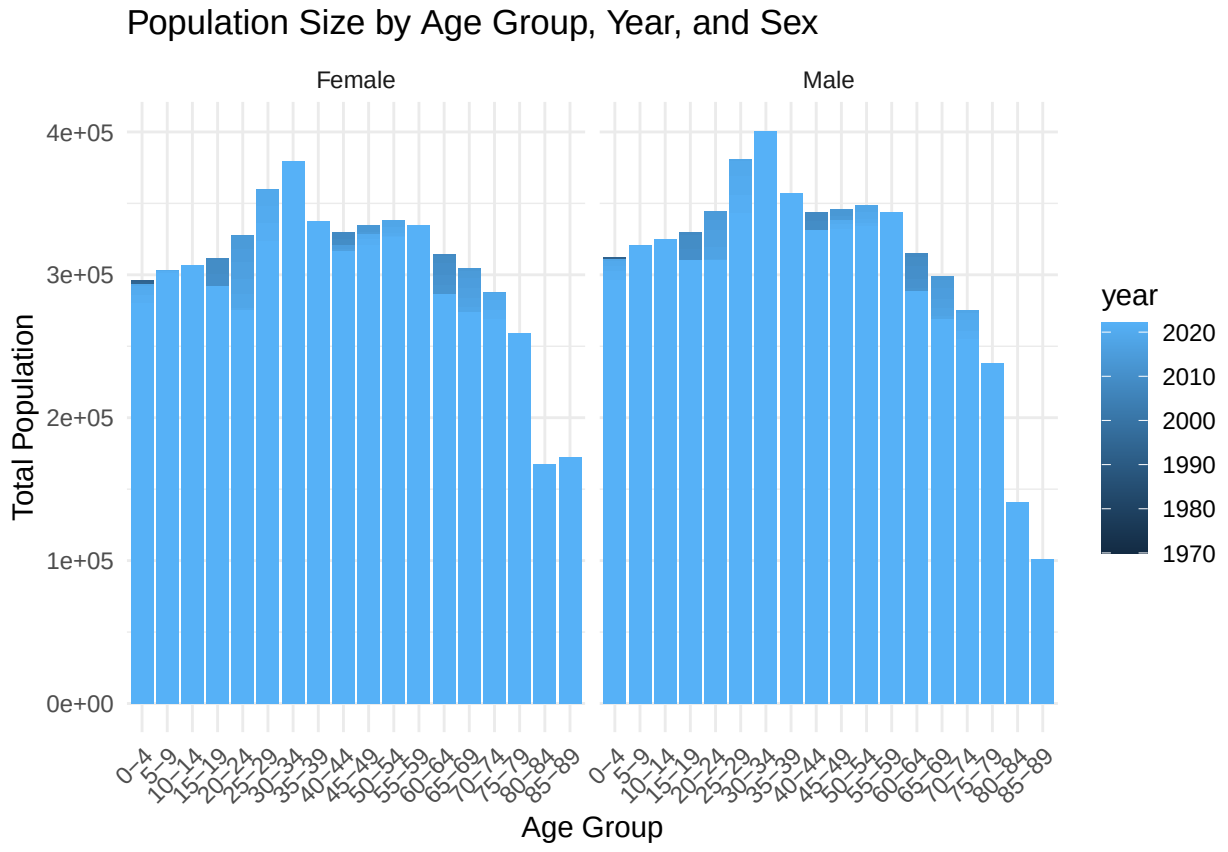
```
# Compare calendar years
all(unique(data$year) %in% unique(population$year))
```

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

```
population_summary <- population %>%
  group_by(agegroup, year, sex) %>%
  summarize(total_population = sum(n_pop, na.rm = TRUE))
```

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

```
ggplot(population_summary, aes(x = agegroup, y = total_population, fill = year)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ sex) +
  labs(title = "Population Size by Age Group, Year, and Sex",
       x = "Age Group",
       y = "Total Population") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#question 4
# Merge the datasets
merged_data <- merge(data, population, by = c("year", "agegroup", "sex"))
# Summarize the total number of cases and population by year and sex
summary_data <- merged_data %>%
  group_by(year, sex) %>%
```

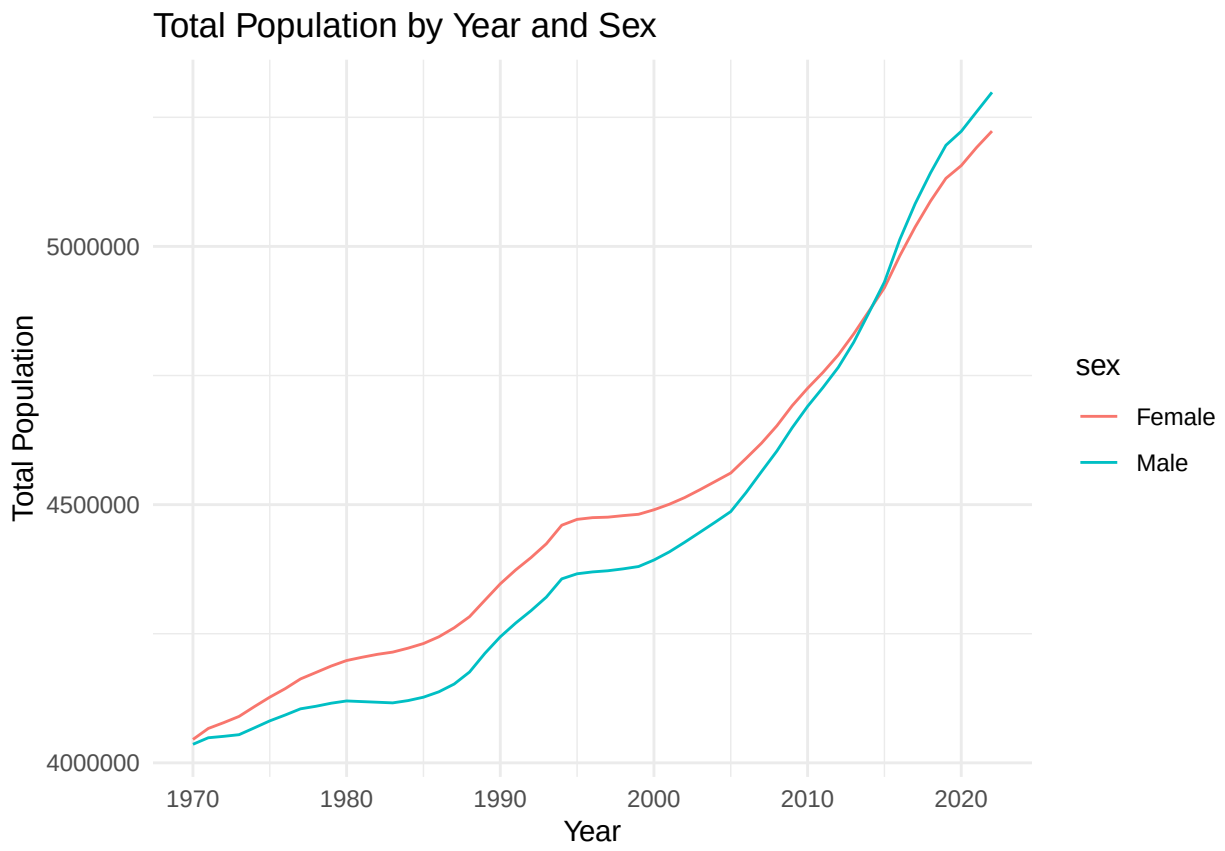
```
  summarize(
    total_cases = sum(n, na.rm = TRUE),
    total_population = sum(n_pop, na.rm = TRUE)
  )
```

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

```
# Check the summarized data
head(summary_data)
```

```
## # A tibble: 6 x 4
## # Groups:   year [3]
##    year sex    total_cases total_population
##   <int> <chr>        <int>            <int>
## 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
```

```
# Plot total population by year and sex
ggplot(summary_data, aes(x = year, y = total_population, color = sex)) +
  geom_line() +
  labs(title = "Total Population by Year and Sex",
       x = "Year",
       y = "Total Population") +
  theme_minimal()
```




```
#question 5
# Add a new variable for incidence rate in the merged data
new_merged_data <- merged_data %>%
  mutate(incidence_rate = n / n_pop)

# Check the updated data
head(new_merged_data)
```

```
##   year agegroup    sex n  n_pop incidence_rate
## 1 1970      0-4 Female 0 280468   0.000000e+00
## 2 1970      0-4   Male 0 296143   0.000000e+00
## 3 1970    10-14 Female 1 257746   3.879789e-06
## 4 1970    10-14   Male 0 272559   0.000000e+00
## 5 1970    15-19 Female 1 269428   3.711567e-06
## 6 1970    15-19   Male 1 281997   3.546137e-06
```

```
# Add a new variable for incidence rate in the summarized data
new_summary_data <- summary_data %>%
  mutate(incidence_rate = total_cases / total_population)

# Check the updated data
head(new_summary_data)
```

```
## # A tibble: 6 x 5
## # Groups:   year [3]
```

```
##    year sex    total_cases total_population incidence_rate
##   <int> <chr>        <int>            <int>          <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
```

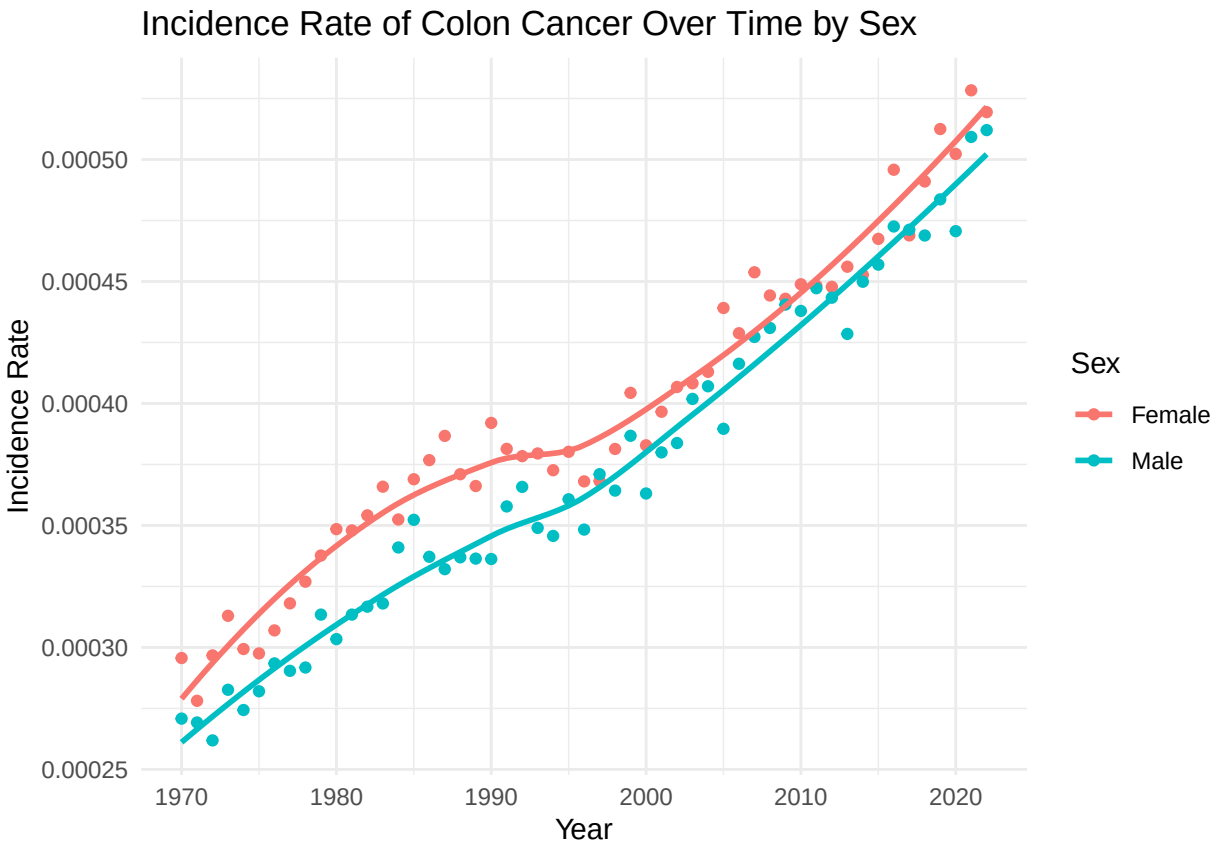
Definition of the incident rate = (total cases /total population) *100

Yes, it looks appropriate.

The incidence rate of colon cancer has steadily increased over calendar time for both males and females. Females consistently have a higher incidence rate of colon cancer compared to males throughout the period.

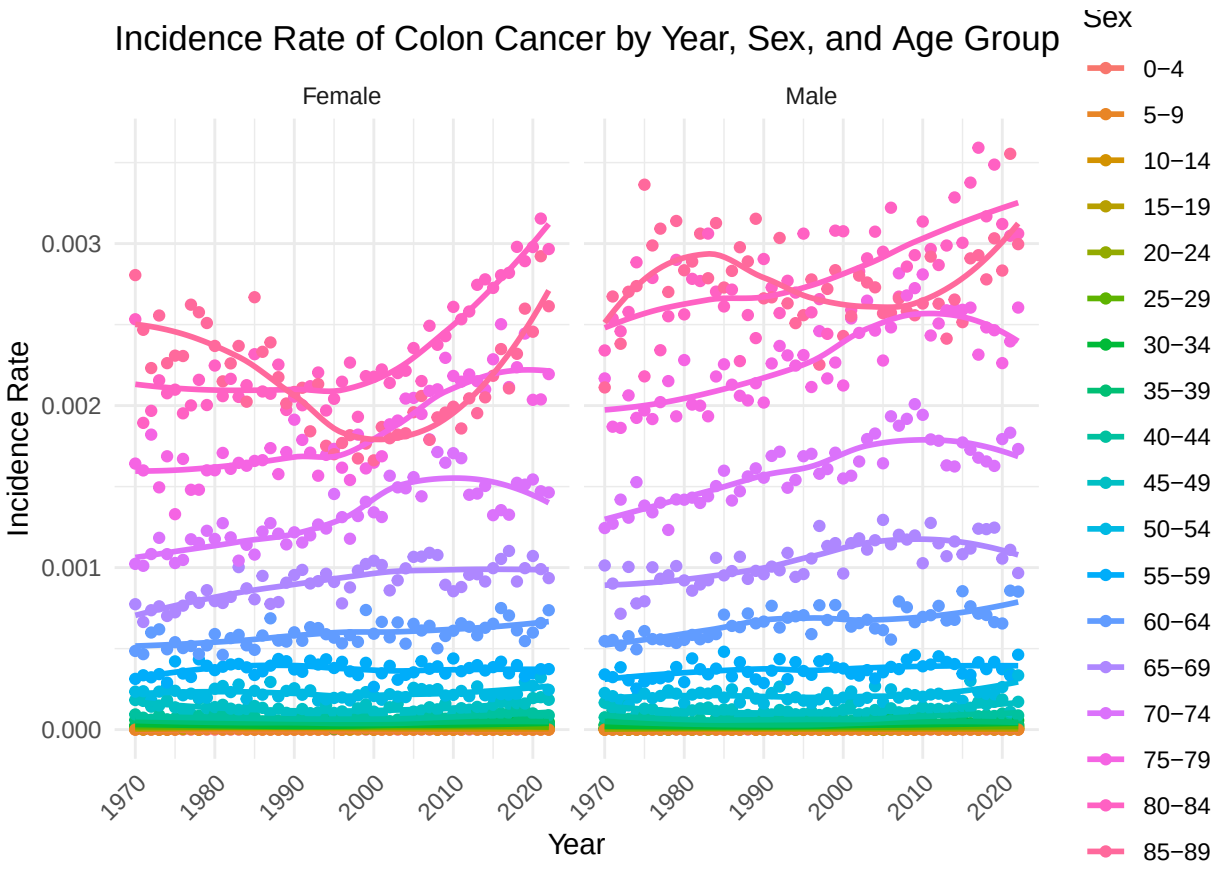
```
#question 6
# Plot incidence rates over calendar time, with smoothers
ggplot(new_summary_data, aes(x = year, y = incidence_rate, color = sex)) +
  geom_point() +  # Points to show individual rates
  geom_smooth(method = "loess", se = FALSE) +  # Apply a smoother
  labs(title = "Incidence Rate of Colon Cancer Over Time by Sex",
       x = "Year",
       y = "Incidence Rate",
       color = "Sex") +
  theme_minimal()
```

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




```
# Plot incidence rates by year, sex, and age group
ggplot(new_merged_data, aes(x = year, y = incidence_rate, color = agegroup)) +
  geom_point() +  # Points for each observation
  geom_smooth(method = "loess", se = FALSE) +  # Apply smoothers
  facet_wrap(~ sex) +  # Separate plots for each age group
  labs(title = "Incidence Rate of Colon Cancer by Year, Sex, and Age Group",
       x = "Year",
       y = "Incidence Rate",
       color = "Sex") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

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



The Poisson model estimates the total number of cases as the dependent variable, using population size as an offset and calendar year and sex as independent variables.

The estimated intercept is -29.69, representing the baseline log-incidence rate when the year is 0 and the reference category for sex (likely Female). The coefficient for the variable year is 0.01094 ($p < 0.001$), indicating a 1.1% annual increase in the incidence rate ((e^(0.01094) - 1) * 100), holding sex constant. The coefficient for sex (Male) is -0.05592 ($p < 0.001$), implying a 5.6% lower incidence rate for males compared to females, holding year constant.

```r
#question 7 assume: main effect model the effect of exposure is the same in all levels of another varia
poisson_model <- glm(
  total_cases ~ year + sex,
  offset = log(total_population),
  family = poisson(link = "log"),
  data = new_summary_data
)

summary(poisson_model)
```

```
##
## Call:
## glm(formula = total_cases ~ year + sex, family = poisson(link = "log"),
##     data = new_summary_data, offset = log(total_population))
##
## 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
```

Based on the model output, the estimated incidence rates in 1970 are 0.0002933 for females and 0.0002774 for males. In 2020, the incidence rates are 0.0005070 for females and 0.0004794 for males. The model assumes that the log-incidence rate changes linearly with the calendar year, leading to an annual incidence rate increase of 1.1% ((e^(0.01094) - 1) * 100).

The key assumptions made in this model:

- A linear relationship between year and log-incidence rate: log( ) = _0 + _1(year) + _2(sex).
- A constant difference between males and females: _2 = -0.0559, males have 5.5% lower incidence rates than females.
- No interactions among year, sex, age group.
- Accurate population size as an offset: the population size is accurately measured and representative of the population at risk for both sexes and years.
- Poisson distribution of the dependent variable: the variance in the number of cases increases with the expected number of cases.

```r
#question 8
# Extract model coefficients
coefficients <- coef(poisson_model)
intercept <- coefficients["(Intercept)"]
year_coeff <- coefficients["year"]
sex_coeff <- coefficients["sexMale"]

# Calculate log incidence rate and incidence rate for 1970 and 2020
incidence_1970_male <- exp(intercept + year_coeff * 1970 + sex_coeff)
incidence_1970_female <- exp(intercept + year_coeff * 1970 )

incidence_2020_male <- exp(intercept + year_coeff * 2020 + sex_coeff)
incidence_2020_female <- exp(intercept + year_coeff * 2020)

# Print results
incidence_1970_male
```

```
##  (Intercept)
## 0.0002773751
```

```r
incidence_1970_female
```

```
##  (Intercept)
## 0.0002933271
```

```
incidence_2020_male
```

```
##  (Intercept)
## 0.0004794199
```

```
incidence_2020_female
```

```
##  (Intercept)
## 0.0005069918
```

Colon cancer is more common in older age groups, and the age distribution has shifted over time. To better estimate incidence rates, we refit the Poisson model by including interaction terms: "year * sex," "year * age group," and "sex * age group."

Based on the updated model output, the estimated incidence rates in 1970 for the 70-74 age group are 0.001350 for males and 0.001098 for females. In 2020, the incidence rates for the same age group are 0.001846 for males and 0.001544 for females.

```
#question9
# Fit a Poisson regression model with interaction terms for flexibility
poisson_model_age <- glm(
  n ~ year * sex + year * agegroup + sex * agegroup,
  offset = log(n_pop),
  family = poisson(link = "log"),
  data = new_merged_data  # Ensure merged_data has cases, population, year, sex, and agegroup
)

# View the summary of the model
summary(poisson_model_age)
```

```
##
## Call:
## glm(formula = n ~ year * sex + year * agegroup + sex * agegroup,
##     family = poisson(link = "log"), data = new_merged_data, offset = log(n_pop))
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -6.879e+01  9.627e+01  -0.715   0.4749
## year                   2.618e-02  4.807e-02   0.545   0.5860
## sexMale                1.086e+00  1.545e+00   0.703   0.4819
## agegroup5-9            5.897e+01  9.865e+01   0.598   0.5500
## agegroup10-14          7.020e+01  9.684e+01   0.725   0.4685
## agegroup15-19          3.521e+01  9.654e+01   0.365   0.7153
## agegroup20-24          4.121e+01  9.648e+01   0.427   0.6693
## agegroup25-29          4.711e+01  9.643e+01   0.489   0.6252
## agegroup30-34          3.562e+01  9.637e+01   0.370   0.7117
## agegroup35-39          4.877e+01  9.634e+01   0.506   0.6127
## agegroup40-44          4.329e+01  9.631e+01   0.449   0.6531
## agegroup45-49          5.194e+01  9.629e+01   0.539   0.5897
## agegroup50-54          5.354e+01  9.628e+01   0.556   0.5782
## agegroup55-59          5.598e+01  9.628e+01   0.581   0.5609
## agegroup60-64          4.976e+01  9.628e+01   0.517   0.6053
```

```
## agegroup65-69          4.985e+01  9.627e+01   0.518   0.6046
## agegroup70-74          4.853e+01  9.627e+01   0.504   0.6142
## agegroup75-79          4.823e+01  9.627e+01   0.501   0.6164
## agegroup80-84          4.867e+01  9.627e+01   0.505   0.6132
## agegroup85-89          6.130e+01  9.627e+01   0.637   0.5243
## year:sexMale          -5.696e-04  3.104e-04  -1.835   0.0665 .
## year:agegroup5-9      -2.792e-02  4.927e-02  -0.567   0.5710
## year:agegroup10-14    -3.283e-02  4.836e-02  -0.679   0.4972
## year:agegroup15-19    -1.496e-02  4.821e-02  -0.310   0.7563
## year:agegroup20-24    -1.788e-02  4.818e-02  -0.371   0.7106
## year:agegroup25-29    -2.080e-02  4.815e-02  -0.432   0.6658
## year:agegroup30-34    -1.484e-02  4.812e-02  -0.308   0.7578
## year:agegroup35-39    -2.117e-02  4.811e-02  -0.440   0.6599
## year:agegroup40-44    -1.816e-02  4.809e-02  -0.378   0.7056
## year:agegroup45-49    -2.220e-02  4.808e-02  -0.462   0.6443
## year:agegroup50-54    -2.274e-02  4.808e-02  -0.473   0.6362
## year:agegroup55-59    -2.372e-02  4.808e-02  -0.493   0.6217
## year:agegroup60-64    -2.037e-02  4.807e-02  -0.424   0.6717
## year:agegroup65-69    -2.020e-02  4.807e-02  -0.420   0.6743
## year:agegroup70-74    -1.935e-02  4.807e-02  -0.403   0.6873
## year:agegroup75-79    -1.904e-02  4.807e-02  -0.396   0.6921
## year:agegroup80-84    -1.914e-02  4.807e-02  -0.398   0.6905
## year:agegroup85-89    -2.552e-02  4.807e-02  -0.531   0.5956
## sexMale:agegroup5-9   -7.823e-01  1.460e+00  -0.536   0.5921
## sexMale:agegroup10-14 -6.568e-01  1.424e+00  -0.461   0.6447
## sexMale:agegroup15-19 -4.856e-01  1.419e+00  -0.342   0.7322
## sexMale:agegroup20-24 -4.125e-01  1.418e+00  -0.291   0.7711
## sexMale:agegroup25-29 -1.100e-01  1.417e+00  -0.078   0.9381
## sexMale:agegroup30-34 -3.506e-02  1.416e+00  -0.025   0.9802
## sexMale:agegroup35-39 -7.101e-02  1.415e+00  -0.050   0.9600
## sexMale:agegroup40-44  1.914e-02  1.415e+00   0.014   0.9892
## sexMale:agegroup45-49 -5.963e-02  1.415e+00  -0.042   0.9664
## sexMale:agegroup50-54 -4.559e-04  1.414e+00   0.000   0.9997
## sexMale:agegroup55-59  4.903e-02  1.414e+00   0.035   0.9723
## sexMale:agegroup60-64  1.673e-01  1.414e+00   0.118   0.9058
## sexMale:agegroup65-69  1.981e-01  1.414e+00   0.140   0.8886
## sexMale:agegroup70-74  2.426e-01  1.414e+00   0.172   0.8638
## sexMale:agegroup75-79  2.643e-01  1.414e+00   0.187   0.8517
## sexMale:agegroup80-84  2.516e-01  1.414e+00   0.178   0.8588
## sexMale:agegroup85-89  3.263e-01  1.414e+00   0.231   0.8175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 406971.1  on 1907  degrees of freedom
## Residual deviance:   2749.5  on 1853  degrees of freedom
## AIC: 11967
##
## Number of Fisher Scoring iterations: 6
```

```r
# Extract model coefficients
coefficients <- coef(poisson_model_age)
```

```r
# males
rate_1970_male <- exp(
  #intercept
  coefficients["(Intercept)"] +
    #year
    coefficients["year"] * 1970 +
    #male
    coefficients["sexMale"] * 1 +
    #age 70-74
    coefficients["agegroup70-74"] * 1 +
    # year*sex
    coefficients["year:sexMale"] * 1970 * 1 +
    # year*age 70-74
    coefficients["year:agegroup70-74"] * 1970 * 1 +
    # sex*age 70-74
    coefficients["sexMale:agegroup70-74"] * 1 * 1)


rate_1970_female <- exp(
  #intercept
  coefficients["(Intercept)"] +
    #year
    coefficients["year"] * 1970 +
    #female
    coefficients["sexMale"] * 0 +
    #age 70-74
    coefficients["agegroup70-74"] * 1 +
    # year*sex
    coefficients["year:sexMale"] * 1970 * 0 +
    # year*age 70-74
    coefficients["year:agegroup70-74"] * 1970 * 1 +
    # sex*age 70-74
    coefficients["sexMale:agegroup70-74"] * 0 * 1)


rate_2020_male <- exp(
  #intercept
  coefficients["(Intercept)"] +
  #year
  coefficients["year"] * 2020 +
  #male
  coefficients["sexMale"] * 1 +
  #age 70-74
  coefficients["agegroup70-74"] * 1 +
  # year*sex
  coefficients["year:sexMale"] * 2020 * 1 +
  # year*age 70-74
  coefficients["year:agegroup70-74"] * 2020 * 1 +
  # sex*age 70-74
  coefficients["sexMale:agegroup70-74"] * 1 * 1)


rate_2020_female <- exp(
```

```
  #intercept
  coefficients["(Intercept)"] +
    #year
    coefficients["year"] * 2020 +
    #female
    coefficients["sexMale"] * 0 +
    #age 70-74
    coefficients["agegroup70-74"] * 1 +
    # year*sex
    coefficients["year:sexMale"] * 2020 * 0 +
    # year*age 70-74
    coefficients["year:agegroup70-74"] * 2020 * 1 +
    # sex*age 70-74
    coefficients["sexMale:agegroup70-74"] * 0 * 1)


rate_1970_male
```

```
## (Intercept)
## 0.001349865
```

```
rate_1970_female
```

```
## (Intercept)
## 0.001097564
```

```
rate_2020_male
```

```
## (Intercept)
## 0.001845669
```

```
rate_2020_female
```

```
## (Intercept)
## 0.001544049
```