AssignmentI

Group2

2024-11-19

Introduction

Data Exploration

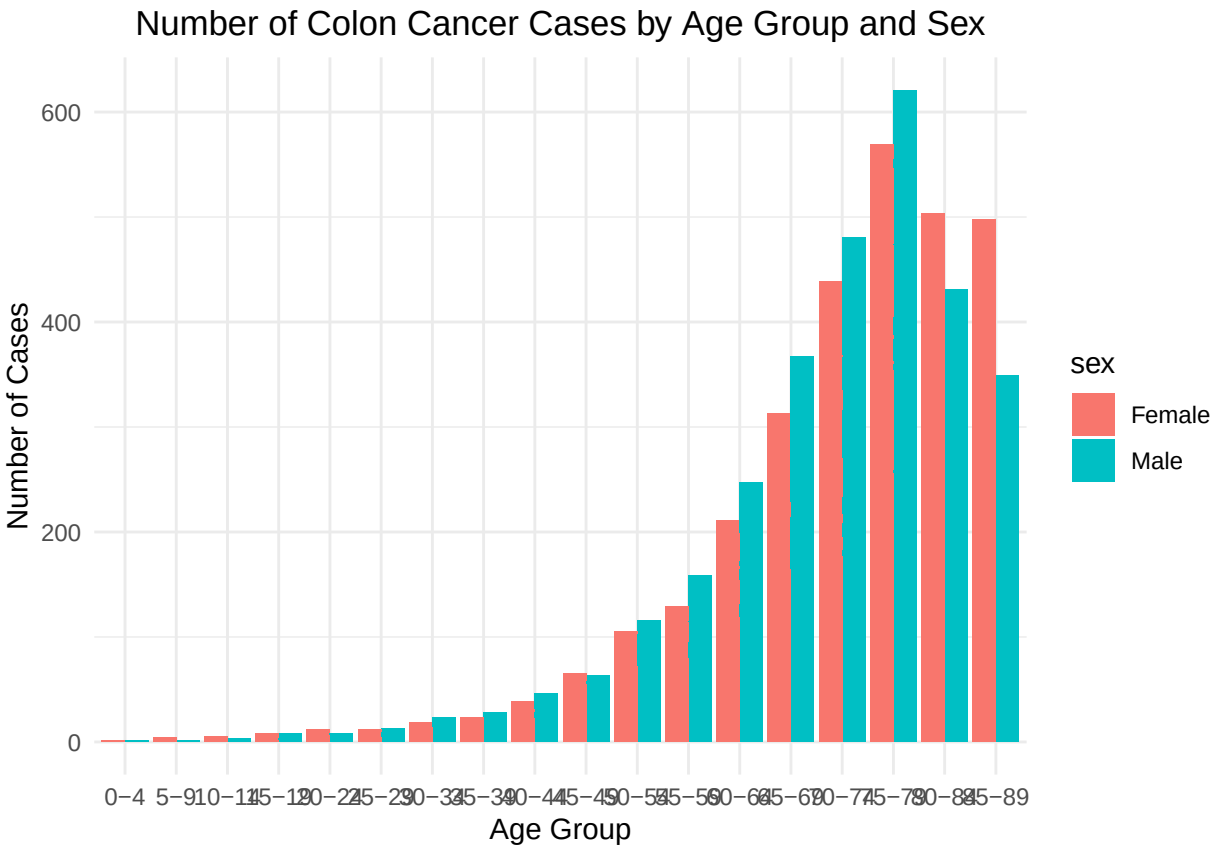
Question 1

Overview :

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

Code :

```
## 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.
```



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.

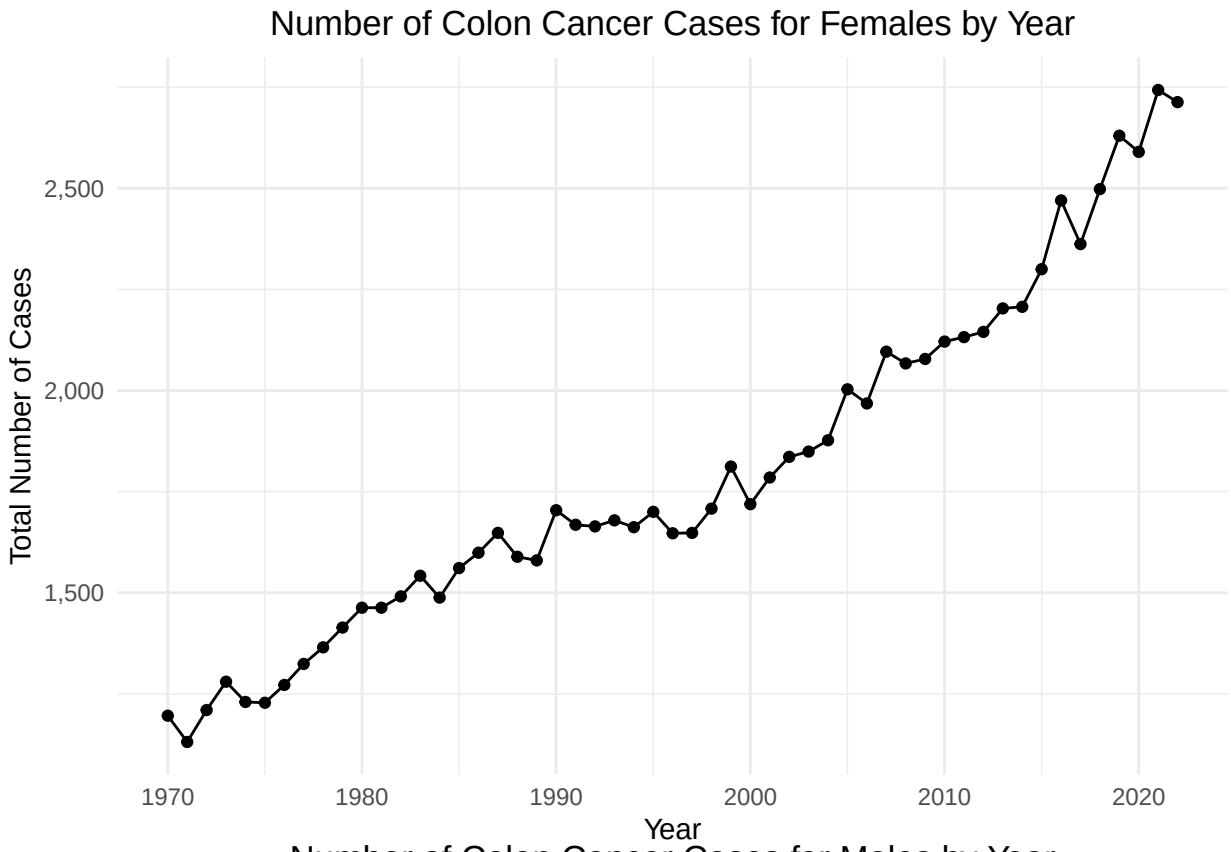
Question 2

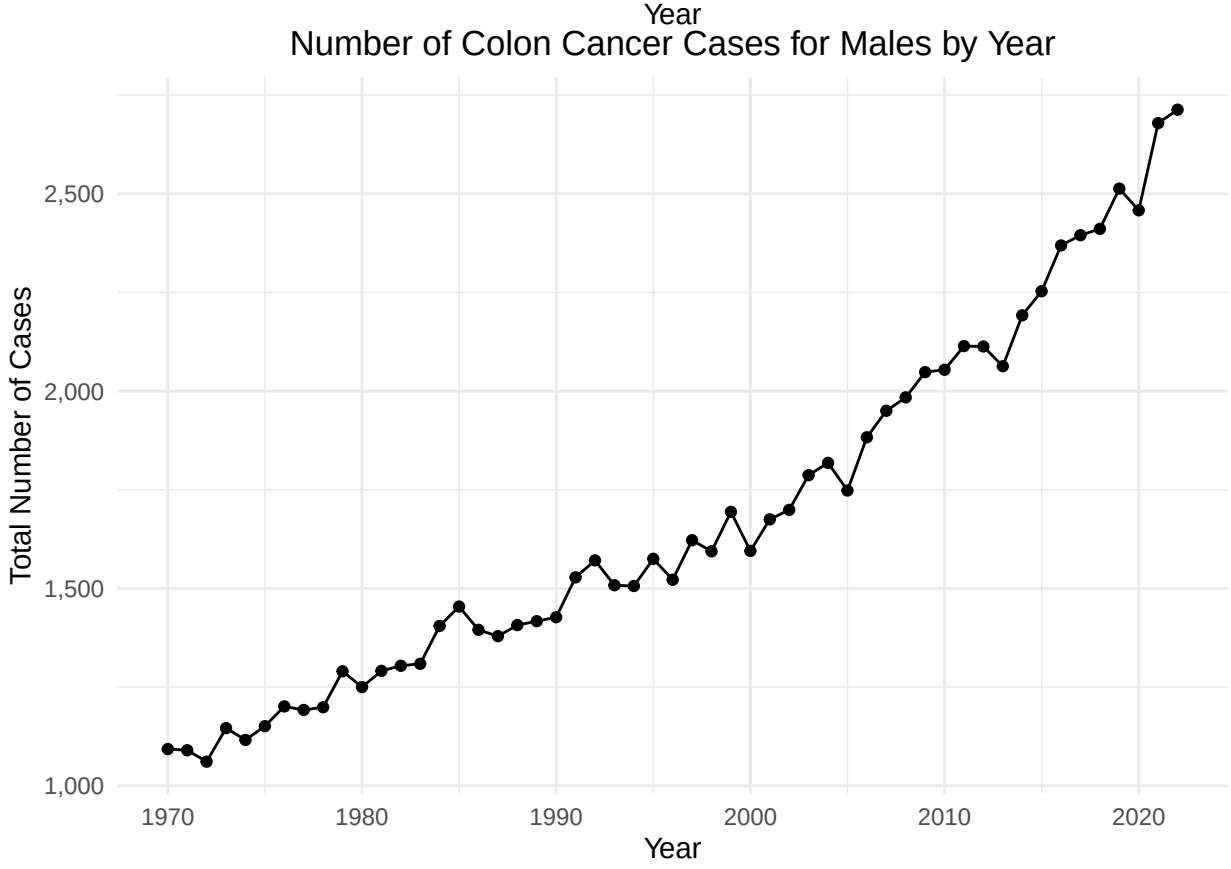
Overview :

Plotting colon cancer cases across subsequent years.

Code :

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


Number of Colon Cancer Cases for Females by Year
Total Number of Cases
2,500
2,000
1,500
1970
1980
1990
2000
2010
2020
Year


Number of Colon Cancer Cases for Males by Year
Total Number of Cases
2,500
2,000
1,500
1,000
1970
1980
1990
2000
2010
2020
Year

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.

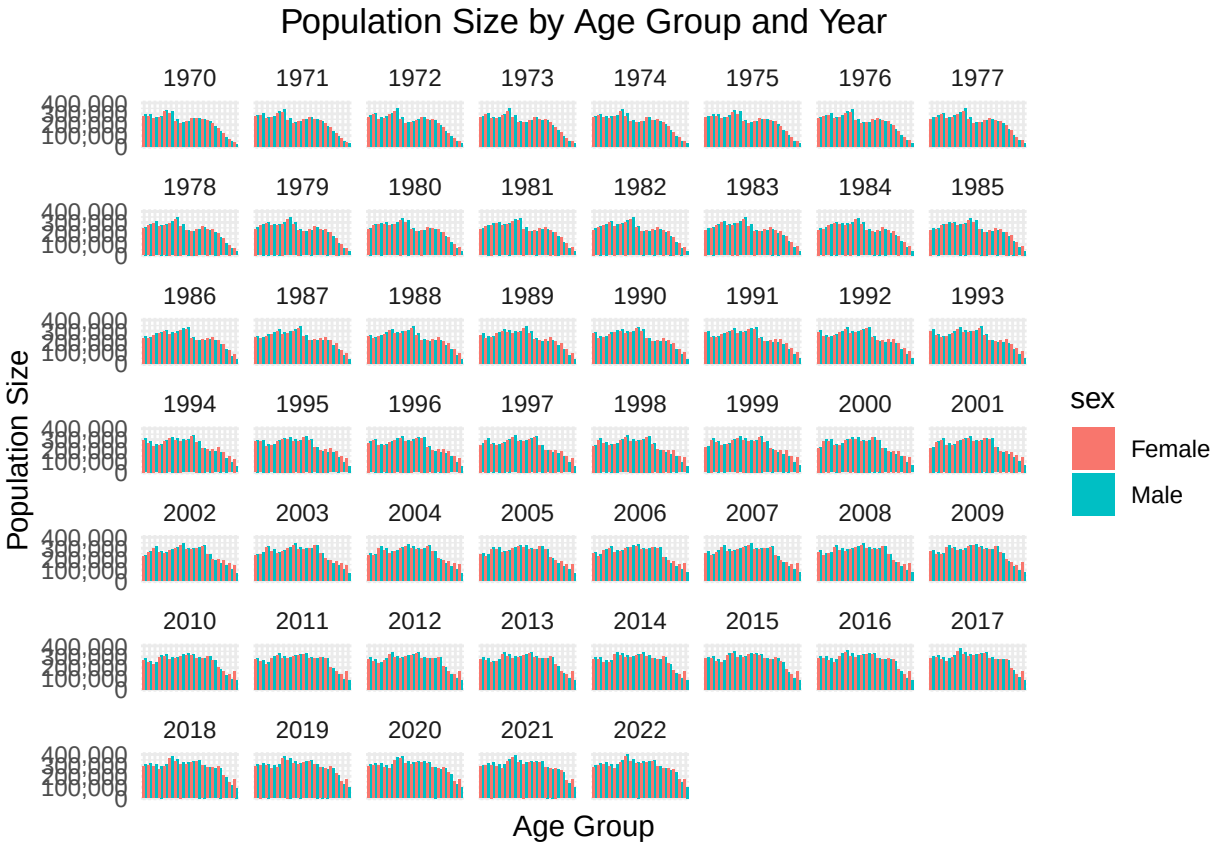
Question 3

Overview :

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

Code :

```
## 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.
```



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 :

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

```
## # 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 :

```
## # 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
```

```
## # 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 :

Code :

Output :

Analysis :

Question 7

Overview :

Code :

Output :

Analysis :

Question 8

Overview :

Code :

Output :

Analysis :

Question 9

Overview :

Code :

Output :

Analysis :

Question 10

Overview :

Code :

Output :

Analysis :

Question 11

Overview :

Code :

Output :

Analysis :

Question 12

Overview :

Code :