Bios 6301: Assignment 2

Yiqing Pan

Due Tuesday, 17 September, 1:00 PM

50 points total.

Add your name as `author` to the file's metadata section.

Submit a single quarto file (named `homework2.qmd`) by email to huiding.chen@vanderbilt.edu. Place your R code in between the appropriate chunks for each question. Check your output by using the `Render` button in RStudio.

1. **Working with data** In the `datasets` folder on the course GitHub repo, you will find a file called `cancer.csv`, which is a dataset in comma-separated values (csv) format. This is a large cancer incidence dataset that summarizes the incidence of different cancers for various subgroups. (18 points)

```
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
```

1. Load the data set into R and make it a data frame called `cancer.df`. (2 points)

```
cancer.df <- read.csv("cancer.csv")
```

2. Determine the number of rows and columns in the data frame. (2)

```
nrow(cancer.df); ncol(cancer.df)
```

```
## [1] 42120
```

```
## [1] 8
```

3. Extract the names of the columns in `cancer.df`. (2)

```
colnames(cancer.df)
```

```
## [1] "year"       "site"       "state"      "sex"        "race"
## [6] "mortality"  "incidence"  "population"
```

4. Report the value of the 3000th row in column 6. (2)

```
cancer.df[3000, 6]
```

```
## [1] 350.69
```

5. Report the contents of the 172nd row. (2)

```
cancer.df[172, ]
```

```
##     year                           site  state  sex  race mortality incidence
## 172 1999 Brain and Other Nervous System nevada Male Black         0         0
##     population
## 172      73172
```

6. Create a new column that is the incidence *rate* (per 100,000) for each row. The incidence rate is the ‘(number of cases)/(population at risk)‘, which in this case means ‘(number of cases)/(population at risk) * 100,000‘. (3)

```
cancer.df <- cancer.df %>% mutate(incidence_rate = incidence/population *100000)

head(cancer.df, 10)
```

```
##    year                           site   state    sex     race mortality
## 1  1999 Brain and Other Nervous System alabama Female    Black      0.00
## 2  1999 Brain and Other Nervous System alabama Female Hispanic      0.00
## 3  1999 Brain and Other Nervous System alabama Female    White     83.67
## 4  1999 Brain and Other Nervous System alabama   Male    Black      0.00
## 5  1999 Brain and Other Nervous System alabama   Male Hispanic      0.00
## 6  1999 Brain and Other Nervous System alabama   Male    White    103.66
## 7  1999 Brain and Other Nervous System  alaska Female    Black      0.00
## 8  1999 Brain and Other Nervous System  alaska Female Hispanic      0.00
## 9  1999 Brain and Other Nervous System  alaska Female    White      0.00
## 10 1999 Brain and Other Nervous System  alaska   Male    Black      0.00
##    incidence population incidence_rate
## 1         19     623475       3.047436
## 2          0      28101       0.000000
## 3        110    1640665       6.704598
## 4         18     539198       3.338291
## 5          0      37082       0.000000
## 6        145    1570643       9.231888
## 7          0      12710       0.000000
## 8          0      11664       0.000000
## 9          0     220036       0.000000
## 10         0      13900       0.000000
```