Breast Cancer Data Analysis

Jagadish Rao

12/18/2021

Introduction

This analysis project is the culmination and final requirement of the **Google Data Analytics Professional Certificate**. It aims to utilize the learning and best practices taught in the course. A sincere 'Thank You' to all my instructors, it was a great, well designed course.

Cancer needs no introduction as a disease that inflicts its pain, suffering, leading to disfigurement and eventual death if not detected and treated early. Unfortunately, the early signs and symptoms of the disease and its progress are often neglected, missed or masked by other diseases. Female breast cancer is a silent killer that can strike anytime, so knowing and proactively acting on the disease patterns is critical.

The identification of breast cancer trends in the USA has been selected for this project. This analysis is approached from a data analytics perspective and has no claims to being a substitute or auxiliary to a professional medical study or advice.

Analysis Objectives

There are many open questions that this analysis seeks to answer:

- Does cancer rate vary by geographical location
- Does race play a part
- Does increasing age lead to a higher risk
- Is there a relationship between race and age as related to cancer risk

Raw Data

United States Cancer Statistics - Incidence: 1999 - 2018, WONDER Online Database. United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; 2021.

Dataset: United States and Puerto Rico Cancer Statistics, 1999-2018 Incidence

https://wonder.cdc.gov/cancer-v2018.HTML

Query Parameters:
Age Groups: 10-14 years; 15-19 years; 20-24 years; 25-29 years; 30-34 years; 35-39 years; 40-44 years; 45-49 years; 50-54 years; 55-59 years; 60-64 years ; 65-69 years; 70-74 years; 75-79 years; 80-84 years; 85+ years
Cancer Sites: Female Breast
Race: American Indian or Alaska Native; Asian or Pacific Islander; Black or African American; White
Sex: Female
States: Alabama (01); Alaska (02); Arizona (04); Arkansas (05); California (06); Colorado (08); Connecticut (09); Delaware (10); District of Columbia (11); Florida (12); Georgia (13); Hawaii (15); Idaho (16); Illinois (17); Indiana (18); Iowa (19); Kansas (20); Kentucky (21); Louisiana (22); Maine (23); Maryland (24); Massachusetts (25); Michigan (26); Minnesota (27); Mississippi (28); Missouri (29); Montana (30); Nebraska (31); Nevada (32); New Hampshire (33); New Jersey (34); New Mexico (35); New York (36); North Carolina (37); North Dakota (38); Ohio (39); Oklahoma (40); Oregon (41); Pennsylvania (42); Rhode Island (44);

South Carolina (45); South Dakota (46); Tennessee (47); Texas (48); Utah (49); Vermont (50); Virginia (51); Washington (53); West Virginia (54); Wisconsin (55); Wyoming (56)
Year: 2000; 2001; 2002; 2003; 2004; 2005; 2006; 2007; 2008; 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018
Group By: Cancer Sites; States; Year; Race; Age Groups
Calculate Rates Per: 100,000
Standard Population: 2000 U.S. Std. Million

Cancer incidence (rate) data is standardized to a population of 100k.
Data is not available if the rate is less than 16.

The raw data was downloaded in four separate text data files due to CDC site download size restrictions. The downloaded files were then renamed to more user friendly titles.

- breast_cancer_2000_2004.txt
- breast_cancer_2005_2009.txt
- breast_cancer_2010_2013.txt
- breast_cancer_2014_2018.txt

Tools

Language: R

Environment: RStudio 2021.09.1+372 "Ghost Orchid" Release (8b9ced188245155642d024aa3630363df611088a, 2021-11-08) for macOS

Data Organization

The downloaded raw data files are tab delimited text files with Header.

The four raw data files are located in a sub folder named "data".

Processing and Analysis

```
library(tidyverse)
library(here)
library(janitor)
library(skimr)
library(validate)
library(usmap)
```

R packages Load

```
bcdata_2000_2004 <- read.delim(here("data", "breast_cancer_2000_2004.txt"))
bcdata_2005_2009 <- read.delim(here("data", "breast_cancer_2005_2009.txt"))
bcdata_2010_2013 <- read.delim(here("data", "breast_cancer_2010_2013.txt"))
bcdata_2014_2018 <- read.delim(here("data", "breast_cancer_2014_2018.txt"))
```

Raw data Read

```
bcdata <- bind_rows(bcdata_2000_2004, bcdata_2005_2009, bcdata_2010_2013, bcdata_2014_2018)
```

Raw data Merge

```
bcdata <- clean_names(bcdata)
```

Column names to lowercase

```
bcdata <- bcdata %>%
  select(states, year, race, age_groups_code, count)
```

Columns selection for analysis

```
bcdata <- bcdata %>%
  rename(state = states, age_grp = age_groups_code, count_100k = count)
```

Columns rename for clarity At this point, all the data has been loaded, relevant columns selected and renamed.

```
skim(bcdata)
```

Data Checking

Table 1: Data summary

Name	bcdata
Number of rows	18745
Number of columns	5
Column type frequency:	
character	3
numeric	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
state	0	1	0	20	280	52	0
race	0	1	0	32	280	5	0
age_grp	0	1	0	5	280	15	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	280	0.99	2009.32	5.45	2000	2005	2009	2014	2018	
count_100k	280	0.99	220.95	320.34	16	42	102	258	3102	

```
tabyl(bcdata, state, year)
```

```
##                 state 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
```

```
##                          0    0    0    0    0    0    0    0    0    0    0
##                Alabama  23   24   23   24   24   23   24   23   24   23   23
##                 Alaska   8    7    5    7    6    8    7    7    8    7    8
##                Arizona  12   13   12   12   12   13   12   13   15   13   17
##               Arkansas   0   17   19   18   18   18   16   18   18   18   20
##             California  38   37   37   40   38   37   38   41   40   40   41
##               Colorado  12   12   13   12   12   12   12   14   13   16   15
##            Connecticut  14   17   19   17   17   18   18   18   18   19   17
##               Delaware  11   10   10   11   13   12   11   13   13   13   13
##  District of Columbia  16   12   14   15   11   11   12   12   13   15   12
##                Florida  27   27   27   25   29   26   28   30   31   30   32
##                Georgia  25   24   25   25   26   25   26   27   27   27   29
##                 Hawaii  17   19   20   20   19   18   18   19   20   19   18
##                  Idaho  11   11   11   11   11   11   11   11   11   11   11
##               Illinois  25   25   25   25   25   25   25   25   25   25   25
##                Indiana  18   22   21   22   21   22   20   22   21   20   20
##                   Iowa  12   12   12   12   12   12   12   12   12   12   12
##                 Kansas  11   11   12   12   12   12   12   12   12   12   13
##               Kentucky  18   18   19   17   17   19   19   18   21   18   19
##              Louisiana  22   23   24   24   24   24   24   23   23   24   23
##                  Maine  11   11   12   11   11   11   11   11   10   11   11
##               Maryland  25   26   25   24   24   26   24   27   29   28   29
##          Massachusetts  16   14   20   18   18   22   18   23   24   22   24
##               Michigan  24   25   25   25   25   25   24   25   25   23   24
##              Minnesota  12   12   12   12   12   12   12   12   12   12   12
##            Mississippi   0    0    0   22   22   23   22   22   23   22   23
##               Missouri  21   22   23   22   23   23   22   24   24   23   22
##                Montana  11   10   10   11   10   10   11   10   11   11   11
##               Nebraska  11   11   11   11   11   11   11   12   11   12   11
##                 Nevada  11   12   11   11   13   12   15   16   13   18   18
##          New Hampshire  11   11   11   11   11   11   11   11   11   11   11
##             New Jersey  30   30   29   31   33   30   32   32   34   33   32
##             New Mexico  11   11   11   11   11   11   11   12   11   12   11
##               New York  33   35   35   36   36   35   35   36   34   35   37
##         North Carolina  25   25   25   25   25   25   25   25   25   26   25
##           North Dakota  10    9   10   10   10   10   10   10   10   10   10
##                   Ohio  25   25   25   24   24   25   24   24   24   24   25
##               Oklahoma  18   20   22   24   21   21   23   22   26   25   23
##                 Oregon  12   12   12   12   12   12   12   12   12   12   12
##           Pennsylvania  25   26   24   24   26   25   26   26   26   26   26
##           Rhode Island  11   11   11   11   11   10   11   11   11   10   11
##         South Carolina  24   24   23   24   24   24   24   23   24   24   24
##           South Dakota   0   11   10   10   10   10   10   10   11   11   10
##              Tennessee  23   22   22   24   25   23   25   24   23   24   23
##                  Texas  30   30   30   31   33   31   33   33   33   33   36
##                   Utah  11   12   11   11   12   11   12   12   12   12   12
##                Vermont  10   11   10   10   11   10   11   10   10   10   10
##               Virginia  26   26   26   26   30   27   28   28   29   29   29
##             Washington  15   17   19   18   19   20   20   20   23   20   22
##          West Virginia  12   11   12   11   11   11   11   11   12   11   11
##              Wisconsin  13   14   14   16   14   15   16   15   16   19   18
##                Wyoming  10   10   10    9    9    9    9    9   10    8    8
##  2011 2012 2013 2014 2015 2016 2017 2018 NA_
##     0    0    0    0    0    0    0    0 280
```

```
##    24   24   25   24   24   24   25   23   0
##     9    9    6    9    9    8    9    8   0
##    17   14   18   21   20   19   21   19   0
##    18   20   21   19   19   20   20   20   0
##    41   42   41   45   41   41   41   43   0
##    12   16   16   14   15   16   17   17   0
##    19   19   20   19   20   18   21   21   0
##    13   12   14   16   15   16   15   16   0
##    14   13   13   13   13   13   15   12   0
##    33   33   33   34   33   35   34   34   0
##    30   27   27   32   32   32   33   34   0
##    20   20   19   19   18   17   19   20   0
##    11   11   12   12   11   11   12   11   0
##    25   25   25   25   25   26   25   25   0
##    22   22   21   22   23   22   21   22   0
##    12   12   12   12   12   12   12   12   0
##    15   14   12   12   13   13   15   13   0
##    19   19   18   19   20   19   20   19   0
##    24   23   23   24   24   25   24   24   0
##    11   11   11   11   11   11   10   11   0
##    29   30   29   30   30   33   33   31   0
##    25   23   29   28   28   29   28   29   0
##    25   26   29   26   27   29   30   30   0
##    15   14   15   16   17   15   19   16   0
##    23   22   23   22   22   22   23   22   0
##    22   23   22   24   23   23   23   23   0
##    10   11   11   10   10   10   11   11   0
##    11   11   11   12   12   12   12   12   0
##    14   18   19   19   19   22   24    0   0
##    11   11   11   11   11   11   11   11   0
##    31   32   34   32   34   34   35   34   0
##    11   12   12   11   11   11   11   13   0
##    37   37   38   38   37   37   37   38   0
##    26   26   26   27   27   28   28   30   0
##    10   10   10   10    9   10   10    9   0
##    24   26   25   25   25   26   28   26   0
##    22   25   26   23   24   25   26   26   0
##    13   13   13   14   16   13   12   13   0
##    28   27   29   30   31   30   31   31   0
##    11   11   10   11   10   11   11   11   0
##    24   24   23   24   24   25   23   24   0
##    10   11   11   10   11   10   10   10   0
##    25   24   23   24   24   23   25   24   0
##    33   36   35   36   35   37   36   38   0
##    12   12   12   12   12   13   12   12   0
##    10   10   10   10   10   10   10   10   0
##    33   30   30   33   32   32   33   31   0
##    25   25   24   27   25   25   26   24   0
##    12   11   11   11   11   11   11   11   0
##    18   19   19   19   19   20   20   18   0
##    10    9   10   10    8    9   10   10   0
```

```
tabyl(bcdata, state, age_grp)
```

```
##                state 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64
```

```
##                         0     0     0     0     0     0     0     0     0
##                Alabama  0     2    31    38    38    38    38    38    38
##                 Alaska  0     0     0     0    16    19    20    19    20
##                Arizona  0     6    19    19    22    31    29    29    33
##               Arkansas  0     0    15    20    34    36    36    35    35
##             California  5    32    57    57    57    60    65    66    66
##               Colorado  0     8    19    19    19    26    26    26    24
##            Connecticut  0     0    19    19    33    39    36    37    37
##               Delaware  0     0     0     6    20    27    32    28    29
##  District of Columbia   0     0     0     0    17    26    32    31    33
##                Florida  0    30    38    42    48    52    54    55    48
##                Georgia  0    22    38    39    46    48    50    46    46
##                 Hawaii  0     0     0    16    28    38    38    38    38
##                  Idaho  0     0     3    19    19    19    19    19    19
##               Illinois  0    20    38    38    38    38    38    38    38
##                Indiana  0     7    19    22    37    38    38    38    38
##                   Iowa  0     0    19    19    19    19    19    19    19
##                 Kansas  0     0    17    19    19    21    20    22    21
##               Kentucky  0     3    19    19    27    38    38    38    35
##              Louisiana  0     1    30    38    38    38    38    38    38
##                  Maine  0     0     1    17    19    19    19    19    19
##               Maryland  0     2    35    41    49    51    52    52    48
##          Massachusetts  0     8    19    29    43    51    50    48    45
##               Michigan  0    17    32    38    40    45    43    40    42
##              Minnesota  0     6    19    19    23    24    25    24    24
##            Mississippi  0     0     6    32    32    32    32    32    32
##               Missouri  0     6    20    33    38    38    38    38    38
##                Montana  0     0     0    10    19    19    19    19    19
##               Nebraska  0     0     7    19    19    19    19    19    19
##                 Nevada  0     0    12    18    19    29    32    33    33
##          New Hampshire  0     0     0    19    19    19    19    19    19
##             New Jersey  0    15    32    53    57    57    57    57    57
##             New Mexico  0     0     5    19    19    19    19    20    19
##               New York  0    29    54    57    57    57    57    57    57
##          North Carolina  0    21    38    38    41    42    43    39    41
##            North Dakota  0     0     0     0    16    19    19    19    19
##                    Ohio  0    19    28    38    39    41    40    38    40
##                Oklahoma  0     0    19    19    32    45    52    55    53
##                  Oregon  0     1    19    19    20    22    22    21    20
##            Pennsylvania  0    19    31    38    46    50    47    46    43
##            Rhode Island  0     0     0    15    19    19    19    19    19
##          South Carolina  0     1    34    38    38    38    38    38    38
##            South Dakota  0     0     0     6    18    18    18    18    18
##               Tennessee  0     8    26    38    38    38    38    38    38
##                   Texas  1    29    43    55    57    57    57    58    55
##                    Utah  0     1    15    19    19    19    19    19    19
##                 Vermont  0     0     0     3    19    19    19    19    19
##                Virginia  0     9    36    41    52    55    55    54    51
##              Washington  0     9    19    26    38    47    47    47    45
##           West Virginia  0     0     4    19    19    19    19    19    19
##               Wisconsin  0     4    19    20    26    36    36    34    34
##                 Wyoming  0     0     0     0    15    19    19    19    19
##  65-69 70-74 75-79 80-84 85+ emptystring_
##      0     0     0     0   0          280
```

```
##     38    38    38    38  38           0
##     19    16    13     3   0           0
##     28    20    19    19  19           0
##     32    31    23    19  21           0
##     64    62    57    57  57           0
##     23    19    19    19  19           0
##     34    30    27    19  19           0
##     25    23    19    19  19           0
##     33    28    22    16  11           0
##     51    46    41    38  38           0
##     43    40    39    38  38           0
##     38    38    37    28  22           0
##     19    19    19    19  19           0
##     38    38    38    38  38           0
##     38    38    36    30  25           0
##     19    19    19    19  19           0
##     23    19    19    19  19           0
##     36    31    30    20  22           0
##     38    38    38    38  38           0
##     19    19    19    19  19           0
##     46    41    39    38  38           0
##     41    35    28    22  19           0
##     41    40    38    38  38           0
##     19    19    19    19  19           0
##     32    32    32    32  32           0
##     38    38    38    37  32           0
##     19    19    19    19  19           0
##     19    19    19    19  19           0
##     29    25    19    18  18           0
##     19    19    19    19  19           0
##     52    51    47    39  38           0
##     19    19    19    19  19           0
##     57    57    54    49  44           0
##     39    38    38    38  38           0
##     19    19    19    19  19           0
##     39    38    38    38  38           0
##     50    43    32    23  19           0
##     19    19    19    19  19           0
##     43    40    38    38  38           0
##     19    19    19    19  19           0
##     38    38    38    38  38           0
##     18    18    18    18  18           0
##     38    38    38    38  36           0
##     55    49    45    40  38           0
##     19    19    19    19  19           0
##     19    19    19    19  19           0
##     46    44    39    38  38           0
##     37    34    26    20  19           0
##     19    19    19    19  19           0
##     29    25    21    19  19           0
##     19    19    19    18  11           0
```

```
tabyl(bcdata, state, race)
```

```
##                 state American Indian or Alaska Native
```

```
##                                                0
##               Alabama                          0
##                Alaska                          2
##               Arizona                          7
##              Arkansas                          0
##            California                         41
##              Colorado                         0
##           Connecticut                         0
##              Delaware                         0
##  District of Columbia                         0
##               Florida                         0
##               Georgia                         0
##                Hawaii                         0
##                 Idaho                         0
##              Illinois                         0
##               Indiana                         0
##                  Iowa                         0
##                Kansas                         0
##              Kentucky                         0
##             Louisiana                         0
##                 Maine                         0
##              Maryland                         0
##         Massachusetts                         0
##              Michigan                         0
##             Minnesota                         0
##           Mississippi                         0
##              Missouri                         0
##               Montana                         0
##              Nebraska                         0
##                Nevada                         0
##         New Hampshire                         0
##            New Jersey                         0
##            New Mexico                         1
##              New York                         0
##        North Carolina                         1
##          North Dakota                         0
##                  Ohio                         0
##              Oklahoma                       127
##                Oregon                         0
##          Pennsylvania                         0
##          Rhode Island                         0
##        South Carolina                         0
##          South Dakota                         0
##             Tennessee                         0
##                 Texas                         4
##                  Utah                         0
##               Vermont                         0
##              Virginia                         0
##            Washington                         5
##         West Virginia                         0
##             Wisconsin                         0
##               Wyoming                         0
##  Asian or Pacific Islander Black or African American White emptystring_
##                          0                         0     0          280
```

```
##                   0                 221   230             0
##                   0                   0   143             0
##                  14                  38   234             0
##                   0                 124   213             0
##                 239                 230   252             0
##                   7                  23   236             0
##                   2                 119   228             0
##                   0                  51   196             0
##                   0                 172    77             0
##                  95                 239   247             0
##                  55                 240   238             0
##                 205                   0   154             0
##                   0                   0   212             0
##                   0                 229   247             0
##                   0                 169   235             0
##                   0                   0   228             0
##                   0                  12   226             0
##                   0                 125   231             0
##                   0                 221   228             0
##                   0                   0   208             0
##                  77                 225   230             0
##                  67                 135   236             0
##                  25                 222   245             0
##                   3                  22   234             0
##                   0                 181   177             0
##                   0                 198   234             0
##                   0                   0   200             0
##                   0                   0   216             0
##                  40                  35   210             0
##                   0                   0   209             0
##                 147                 222   243             0
##                   0                   0   214             0
##                 201                 238   247             0
##                  16                 230   247             0
##                   0                   0   187             0
##                   9                 218   247             0
##                   0                  87   228             0
##                  10                   0   229             0
##                  49                 221   247             0
##                   0                   0   205             0
##                   0                 224   229             0
##                   0                   0   186             0
##                   0                 214   236             0
##                 149                 238   248             0
##                   0                   0   225             0
##                   0                   0   193             0
##                  95                 226   237             0
##                 137                  35   237             0
##                   0                   0   213             0
##                   0                  90   232             0
##                   0                   0   177             0
```

```
#create data validation rules
rules <- validator(state != "",
```

```
  year >= 2000,
  year <= 2018,
  race != "",
  age_grp != "",
  count_100k >= 0)

# verify data using rules
confront(bcdata, rules)
```

```
## Object of class 'validation'
## Call:
##     confront(dat = bcdata, x = rules)
##
## Rules confronted: 6
##    With fails   : 3
##    With missings: 3
##    Threw warning: 0
##    Threw error  : 0
```

```
# The above data checking shows that:
# Arkansas is missing data for year 2000
# Mississippi is missing data for year 2000-2002
# Nevada is missing data for year 2018
# South Dakota is missing data for year 2000
# age_group 20-24 is missing data for almost all (48) states
# Race - American Indian or Alaska Native is missing data for most states (42), as expected

# For an unbiased analysis -
# Remove data for all states for years 2000-2002, 2018
# Remove data for age_group 20-24
bcdata <- bcdata %>%
  filter(year %in%  c(2003:2017),
         age_grp != "20-24")

# Verify that no more errors
confront(bcdata, rules)
```

```
## Object of class 'validation'
## Call:
##     confront(dat = bcdata, x = rules)
##
## Rules confronted: 6
##    With fails   : 0
##    With missings: 0
##    Threw warning: 0
##    Threw error  : 0
```

```
write_csv(bcdata, here("data", "bcdata.csv"))
```

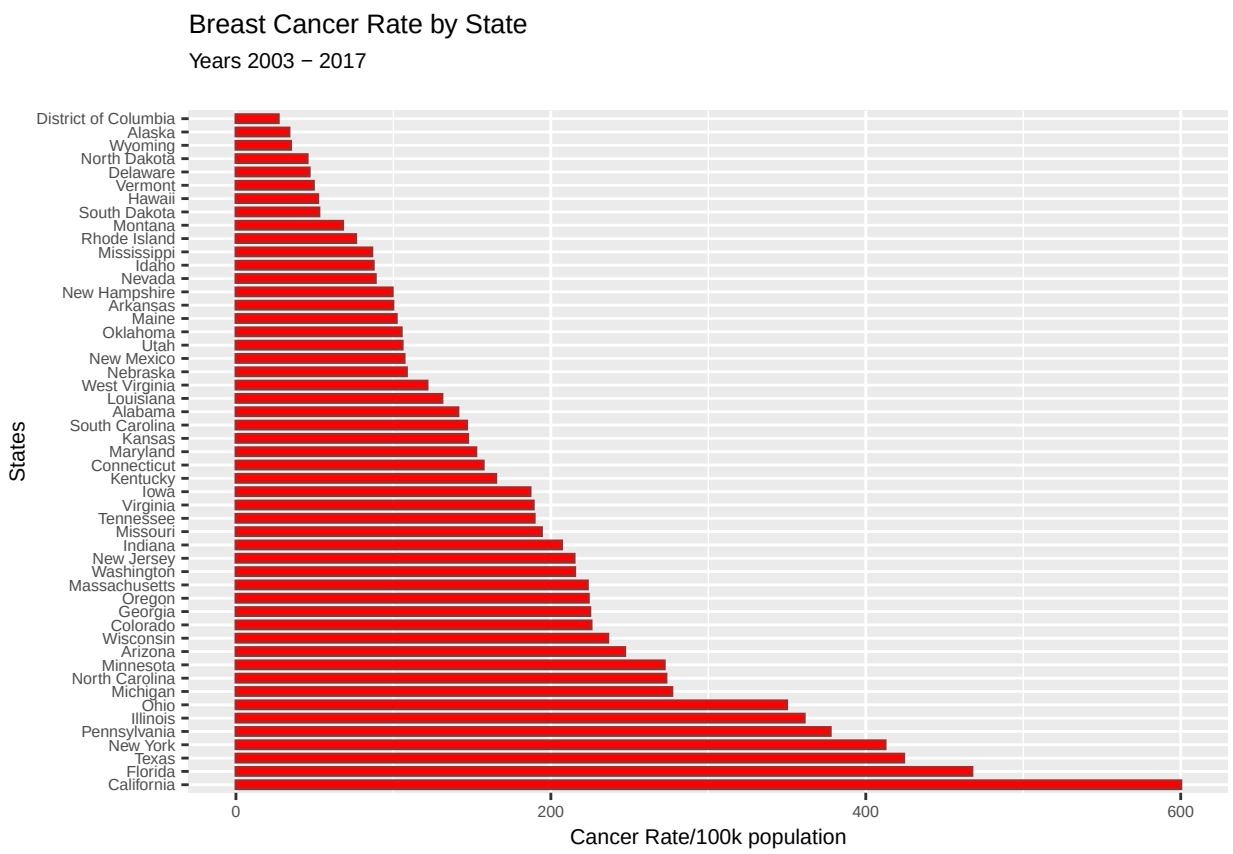
Save the cleaned data set

Data Analysis

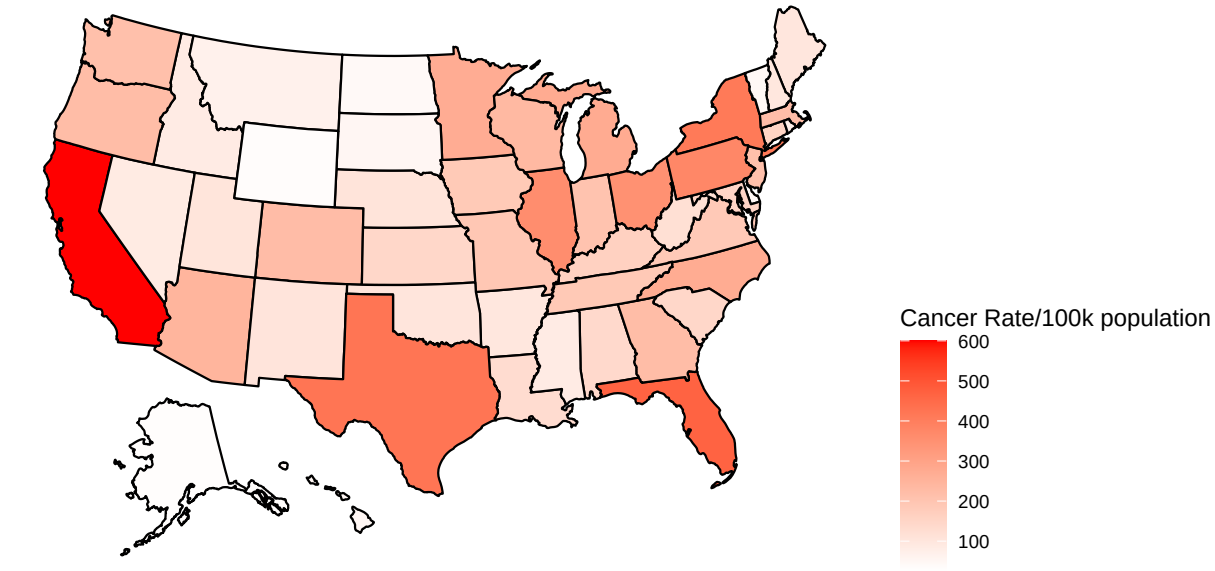
Cancer distribution by State

```
# find the mean cancer rate per state
bcdata_state_mean <-
  bcdata %>%
  group_by(state) %>%
  summarise(state_mean = mean(count_100k))

# plot a column graph
bcdata_state_mean %>%
  ggplot(aes(x = state_mean, y = reorder(state, -state_mean)))  +
  geom_col(color = "brown", fill="red", width = 0.6) +
  theme(text = element_text(size = 8)) +
  xlab("Cancer Rate/100k population") +
  ylab("States") +
  labs(title = "Breast Cancer Rate by State", subtitle = "Years 2003 - 2017\n")
```

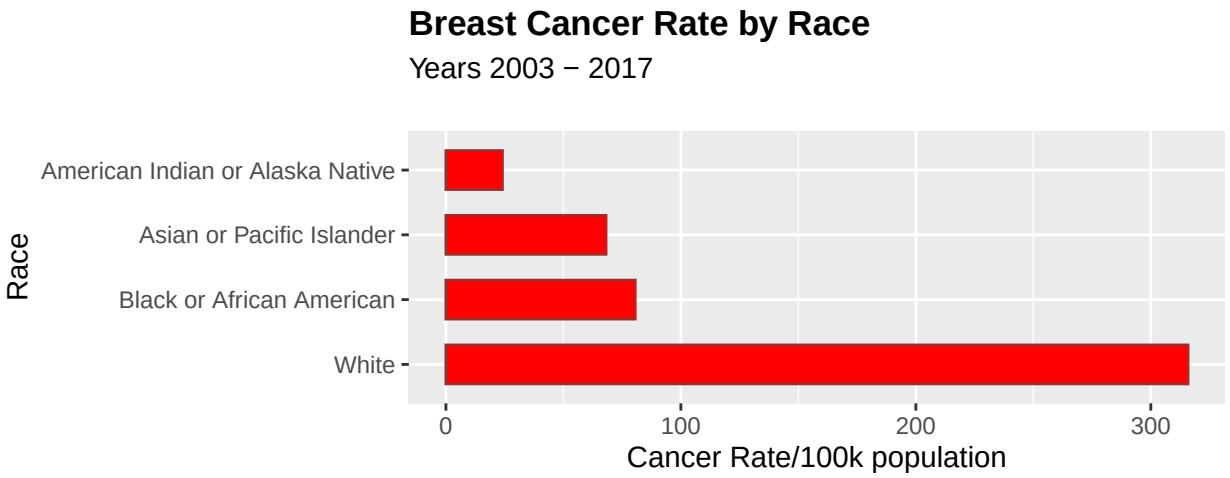



```
# plot a map
plot_usmap(data = bcdata_state_mean, regions = "states", values = "state_mean") +
  scale_fill_continuous(name = "Cancer Rate/100k population", low = "white", high = "red") +
  theme(legend.position = "right") +
  labs(title = "Breast Cancer Rate by State", subtitle = "Years 2003 - 2017") +
  theme(plot.title = element_text(face="bold"))
```



Cancer distribution by Race

```
# plot of cancer rate by race
bcdata %>%
  group_by(race) %>%
  summarise(race_rate = mean(count_100k)) %>%
  ggplot(aes(x = race_rate, y = reorder(race, -race_rate)))  +
  geom_col(color = "brown", fill="red", width = 0.6) +
  theme(aspect.ratio = 1/3)  +
  xlab("Cancer Rate/100k population") +
  ylab("Race") +
  labs(title = "Breast Cancer Rate by Race", subtitle = "Years 2003 - 2017\n") +
  theme(plot.title = element_text(face="bold"))
```




```
# find the mean cancer rate by Race
bcdata_race <-
  bcdata %>%
```

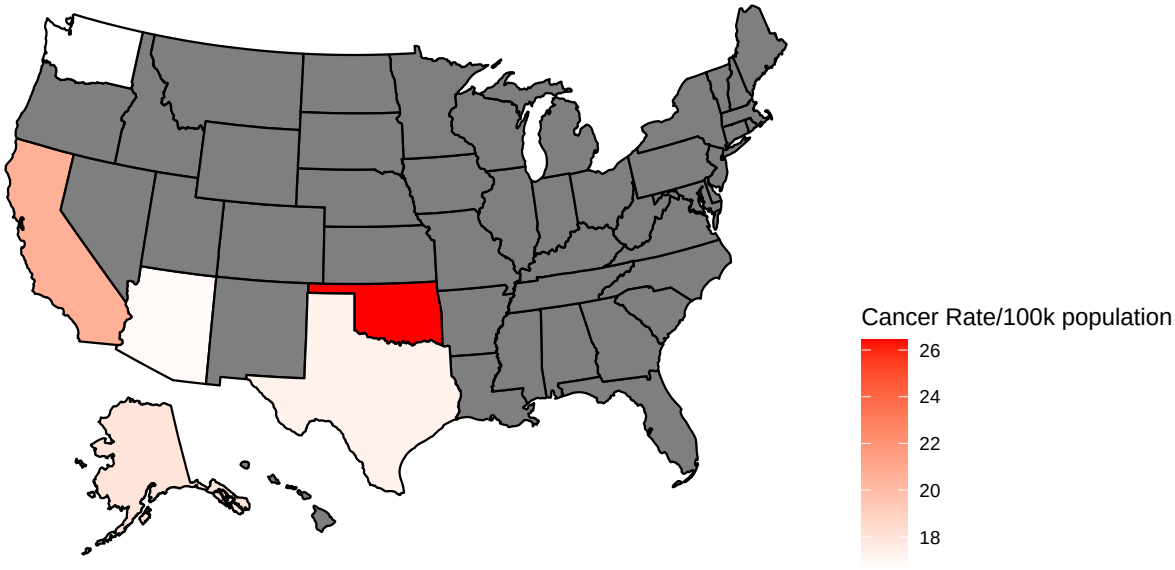
```
  group_by(race, state) %>%
  summarise(race_mean = mean(count_100k))

# plot race = American Indian
bcdata_race_type <- bcdata_race %>%
  filter(race == "American Indian or Alaska Native")

# plot a map
plot_usmap(data = bcdata_race_type, regions = "states", values = "race_mean") +
  scale_fill_continuous(name = "Cancer Rate/100k population", low = "white", high = "red") +
  theme(legend.position = "right") +
  labs(title = "Cancer Rate by State and Race", subtitle = "Race - American Indian or Alaska Native") +
  theme(plot.title = element_text(face="bold"))
```

Cancer Rate by State and Race

Race − American Indian or Alaska Native

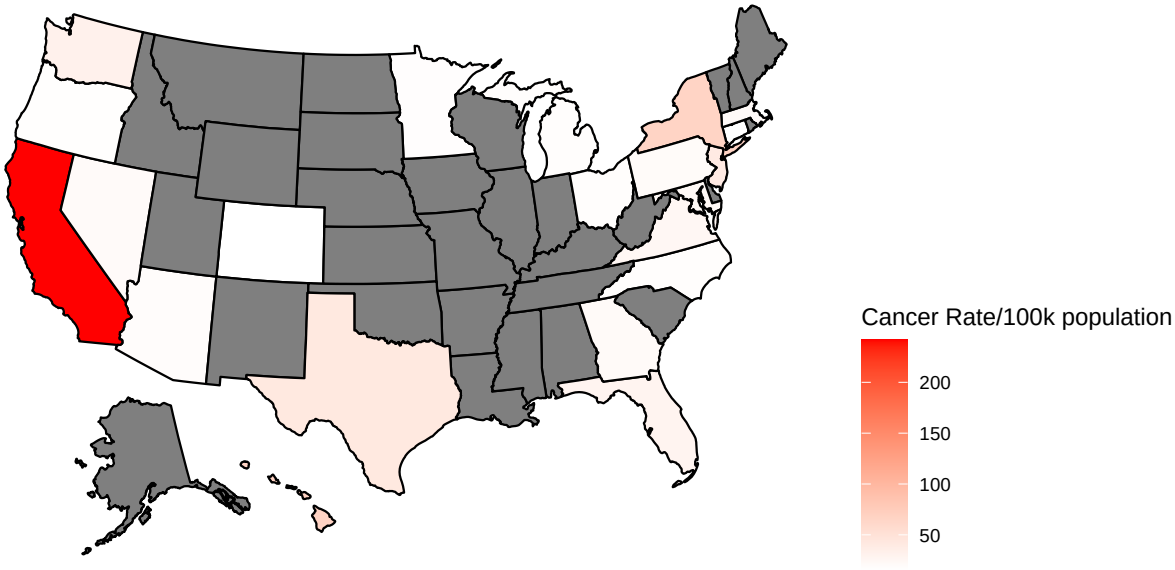



```
# plot race = Asian or Pacific Islander
bcdata_race_type <- bcdata_race %>%
  filter(race == "Asian or Pacific Islander")

# plot a map
plot_usmap(data = bcdata_race_type, regions = "states", values = "race_mean") +
  scale_fill_continuous(name = "Cancer Rate/100k population", low = "white", high = "red") +
  theme(legend.position = "right") +
  labs(title = "Cancer Rate by State and Race", subtitle = "Race - Asian or Pacific Islander") +
  theme(plot.title = element_text(face="bold"))
```

Cancer Rate by State and Race

Race − Asian or Pacific Islander

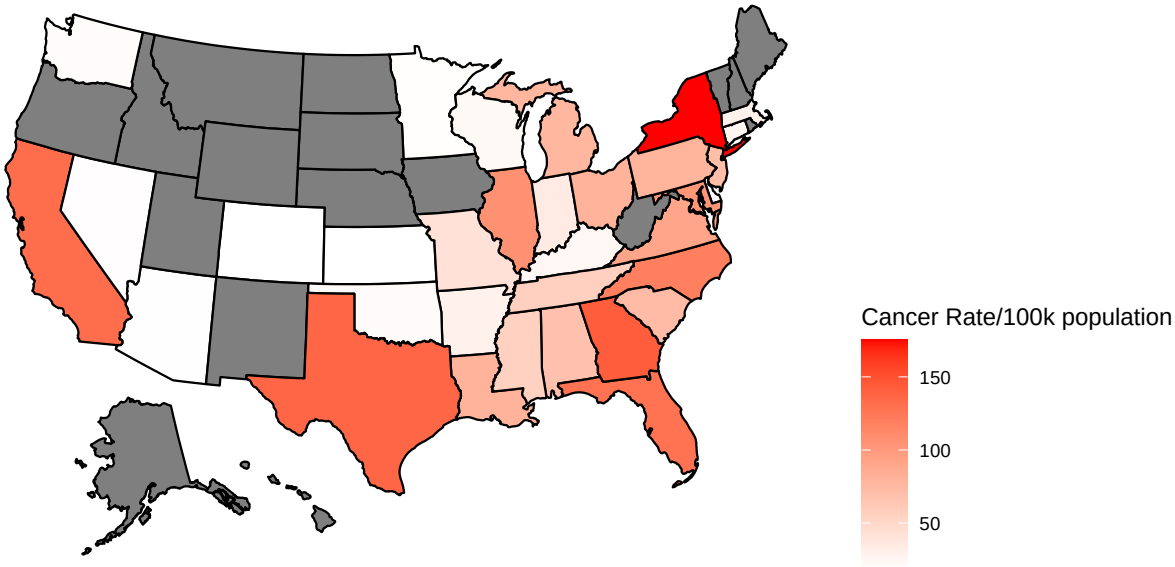



```
# plot race = Black or African American
bcdata_race_type <- bcdata_race %>%
  filter(race == "Black or African American")

# plot a map
plot_usmap(data = bcdata_race_type, regions = "states", values = "race_mean") +
  scale_fill_continuous(name = "Cancer Rate/100k population", low = "white", high = "red") +
  theme(legend.position = "right") +
  labs(title = "Cancer Rate by State and Race", subtitle = "Race - Black or African American") +
  theme(plot.title = element_text(face="bold"))
```

Cancer Rate by State and Race

Race − Black or African American

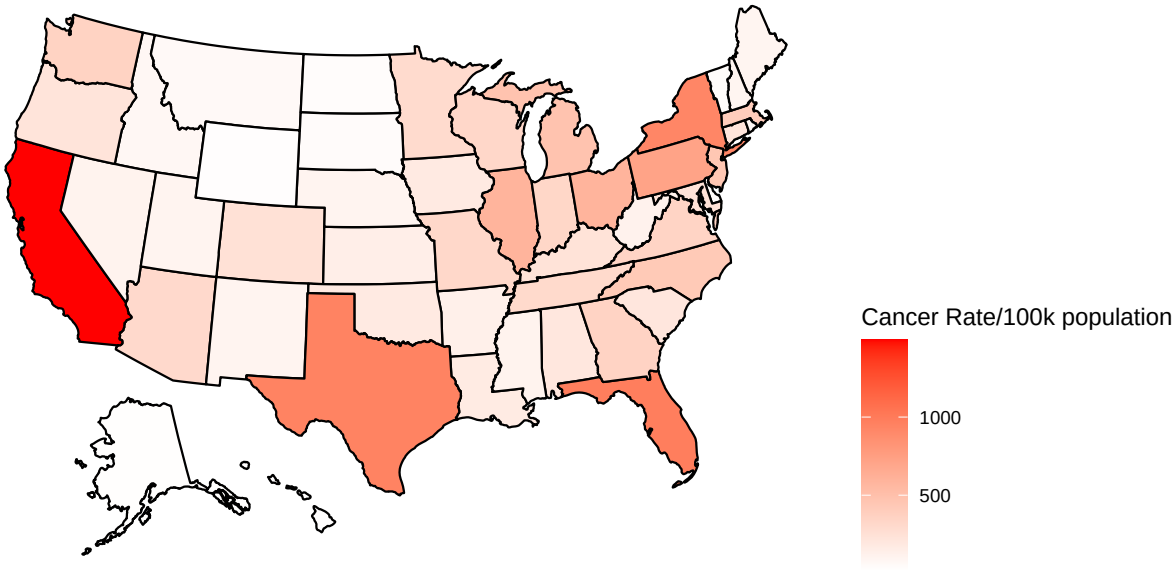


```
# plot race = White
bcdata_race_type <- bcdata_race %>%
  filter(race == "White")

# plot a map
plot_usmap(data = bcdata_race_type, regions = "states", values = "race_mean") +
  scale_fill_continuous(name = "Cancer Rate/100k population", low = "white", high = "red") +
  theme(legend.position = "right") +
  labs(title = "Cancer Rate by State and Race", subtitle = "Race - White") +
  theme(plot.title = element_text(face="bold"))
```

Cancer Rate by State and Race

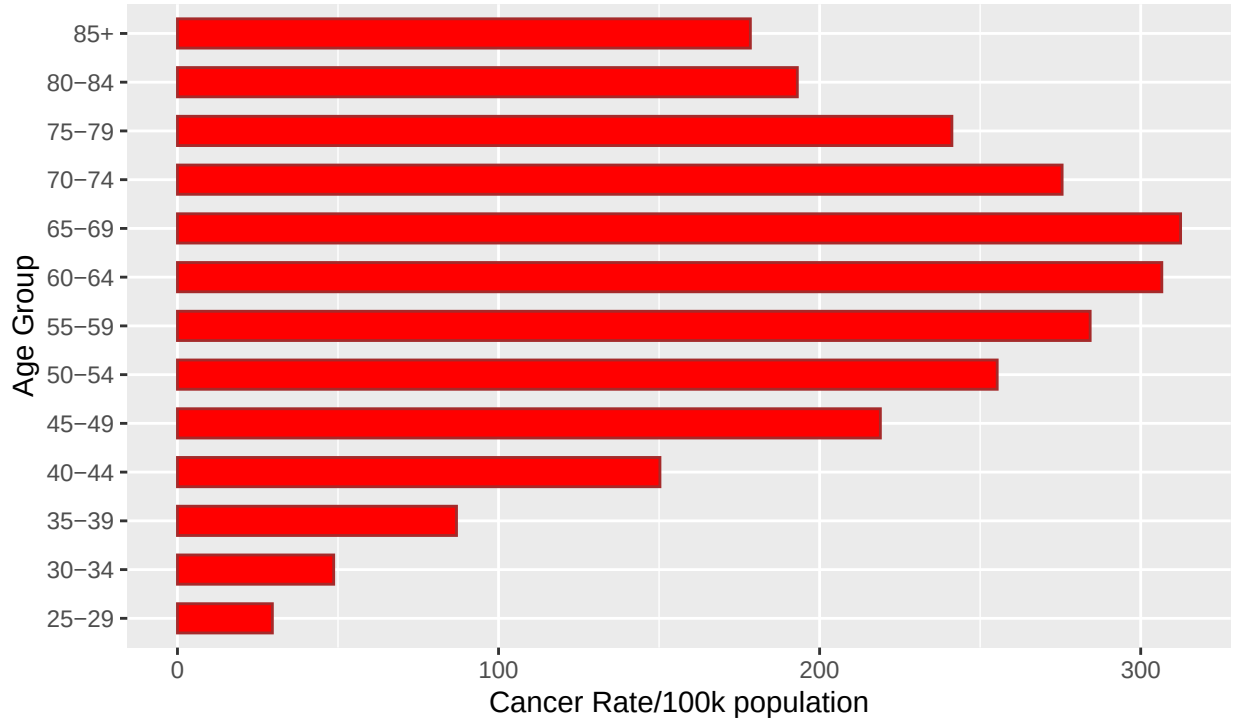
Race − White



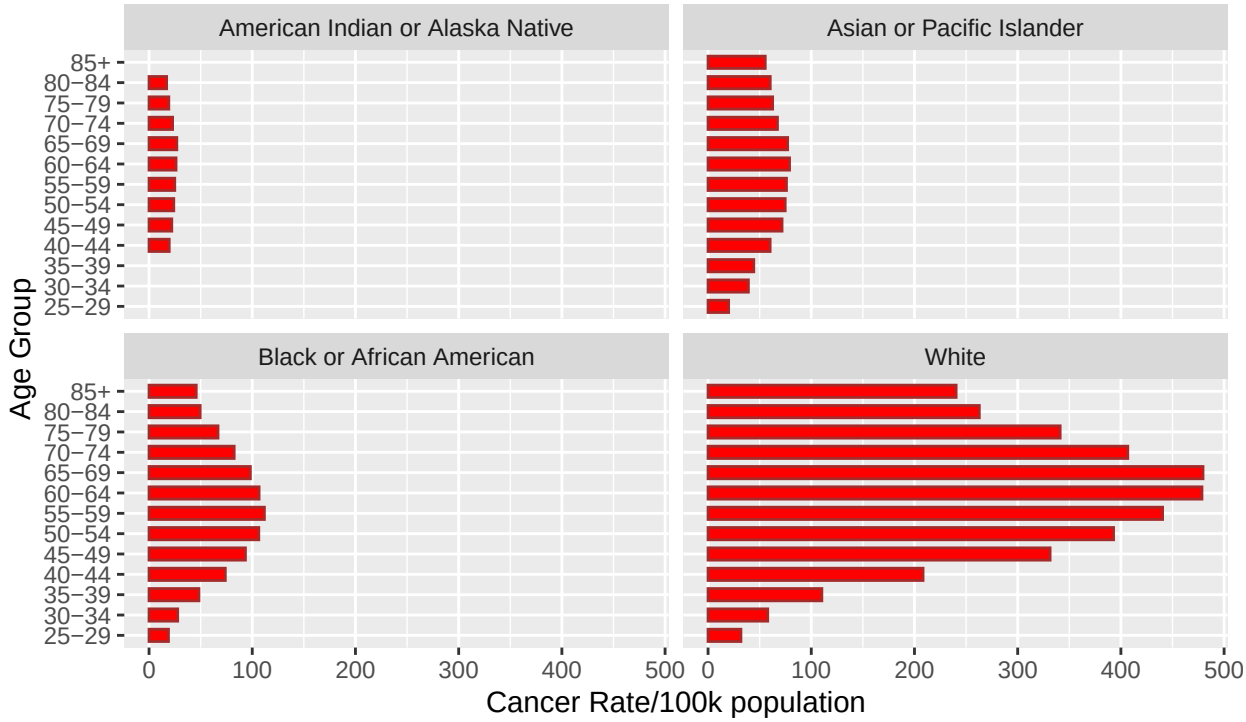
Cancer Distribution by Age Group

```
bcdata_age_grp_mean <-
  bcdata %>%
  group_by(age_grp) %>%
  summarise(age_grp_rate = mean(count_100k))

# plot the graph
bcdata_age_grp_mean  %>%
  ggplot(aes(x = age_grp_rate, y = age_grp))  +
  geom_col(color = "brown", fill="red", width = 0.6) +
  xlab("Cancer Rate/100k population") +
  ylab("Age Group") +
  labs(title = "Breast Cancer Rate by Age Group", subtitle = "Years 2003 - 2017\n") +
  theme(plot.title = element_text(face="bold"))
```




```
bcdata %>%
  group_by(age_grp, race) %>%
  summarise(age_grp_rate = mean(count_100k)) %>%
  ggplot(aes(x = age_grp_rate, y = age_grp))  +
  geom_col(color = "brown", fill="red", width = 0.6) +
  xlab("Cancer Rate/100k population") +
  ylab("Age Group") +
  labs(title = "Breast Cancer Rate by Age Group and Race", subtitle = "Years 2003 - 2017\n") +
  theme(plot.title = element_text(face="bold")) +
  facet_wrap(~race)
```



Conclusions

The following inferences are derived from the analysis:

1. Cancer incidence rate varies by geographical location (State). Since individual states are comprised of populations of different races and ages, it is the composite of these variables that determines the cancer rate of a state. For instance, California and Florida have the highest incidence of cancer rates. Analysis shows us that white women have the highest rate, along with higher rates in late to middle age. Combining these two insights with the tendency of California and Florida to be retirement states is the likely cause that these states have high cancer rates.
2. Race has an impact on the cancer rate with White being the highest. This may be due to factors such as lifestyle and genetics, and it is possible for results to be influenced by additional factors listed below.
3. Increasing age leads to a higher risk, with a peak at 65-69 years followed by decreasing risk. This is observed across all races, however the overall cancer rate varies by race, with White being the highest.
4. There is a relationship between race and age related cancer risk. The pattern of increasing cancer rate as age increases upto 65 years followed by a decline is consistent across all races. However the overall cancer rate varies by race.

Cancer incidence rate variations are also likely due to certain factors not addressed by this analysis.

1. Better access to medical care (Doctor/Hospital to population ratio) that leads to a higher detection/incidence rate
2. Dietary choices
3. Income levels that afford access to medical care or restrict it

4. Awareness and motivation for regular breast cancer screening

To summarise, there are numerous factors, many of which are outside the scope of this analysis that play a part in predicting a breast cancer probability. The variables used in this analysis are historical statistics that only serve as trend indicators and merit deeper scientific scrutiny with relevent datasets.
