A Data Analysis on United States Cancer Statistics (USCS)

Yashar Mansouri

May 1, 2019

Data

-Data Source: https://www.cdc.gov/cancer/uscs/dataviz/download_data.htm

-Description:

1. Data is collected between 1999-2015
2. Contains values from 50 States and different counties
3. 24 Million Cancer Cases
4. Contains different Variables such as Age, Sex, Race, etc.
5. Per Different Cancer sites or All Sites Combined
6. Reported from hospitals, physicians and labs across U.S. to central cancer registries supported by CDC and NCI

Terms

1. Incidence:

"Total number of new cancer cases diagnosed in a specific year in the population category of interest, divided by the at-risk population for that category and multiplied by 100,000 (cancers by primary site)"

2. Mortality:

"Total number of cancer deaths during a specific year in the population category of interest, divided by the at-risk population for that category and multiplied by 100,000"

3. Age Adjusted Rate:

The number of cases (or deaths) per 100,000 people and are age-adjusted to the 2000 U.S. standard population (19 age groups – Census P25–1130)

-Ensures that differences in incidence or deaths from one year to another, or between one geographic area and another, are not due to differences in the age distribution of the populations being compared

Importing the data

```r
library(tidyverse)
library(sf)
library(maps)
library(tmap)
```

```
byarea <- read_delim("../FP/BYAREA.txt", delim = "|")
byareaCounty <- read_delim("../FP/BYAREA_COUNTY.txt", delim = "|")
bysite <- read_delim("../FP/BYSITE.txt", delim = "|")
byage <- read_delim("../FP/BYAGE.txt", delim = "|")
```

```
## Warning: 17214 parsing failures.
##    row  col               expected actual              file
## 220448 YEAR no trailing characters  -2015 '../FP/BYAGE.txt'
## 220449 YEAR no trailing characters  -2015 '../FP/BYAGE.txt'
## 220450 YEAR no trailing characters  -2015 '../FP/BYAGE.txt'
## 220451 YEAR no trailing characters  -2015 '../FP/BYAGE.txt'
## 220452 YEAR no trailing characters  -2015 '../FP/BYAGE.txt'
## ...... .... ...................... ...... .................
## See problems(...) for more details.
```

```
#importing this dataset gives parsing error on "2011-2015" YEAR set,
#yet this is fine since after the import all observations remain the same.

read_csv("../FP/rural.csv")%>%
  select(1, 2, 3, 8)%>%
  slice(1:3142)-> #taking out the last rows
  rural
```

```
## Warning: Missing column names filled in: 'X9' [9], 'X10' [10], 'X11' [11],
## 'X12' [12], 'X13' [13], 'X14' [14]
```

```
names(rural)[4]<-"percent"


#Defining NA values
byarea <- na_if(byarea, '~')
byarea <- na_if(byarea, '-')
byareaCounty <- na_if(byareaCounty, '~')
byareaCounty <- na_if(byareaCounty, '.')
byareaCounty <- na_if(byareaCounty, '-')
bysite <- na_if(bysite, '~')
bysite <- na_if(bysite, '.')
byage <- na_if(byage, '~')
byage <- na_if(byage, '.')


#parsing some numerical values that are in our data exploration
byage%>%
  mutate(COUNT = parse_number(COUNT))%>%
  mutate(RATE = parse_number(RATE)) ->
  byage

bysite%>%
  mutate(AGE_ADJUSTED_RATE = parse_number(AGE_ADJUSTED_RATE)) ->
  bysite
byareaCounty%>%
  mutate(AGE_ADJUSTED_RATE = parse_number(AGE_ADJUSTED_RATE)) ->
  byareaCounty
```

Hypotheses

Our initial hypotheses were as below:

1. Between 2011-2015 the rate of cancer in rural areas should be lower than urban areas.
2. The mortality rate of cancer in elderlies are higher than the other age groups.
3. There's an association between the death rate of skin cancer and different ethnicities.
4. Males are more prone to new cancers than females.
5. Rate of new cancers during the 1999-2015 should increase.

HYPOTHESIS 1

Between 2011-2015 the rate of cancer in rural areas should be lower than urban areas.

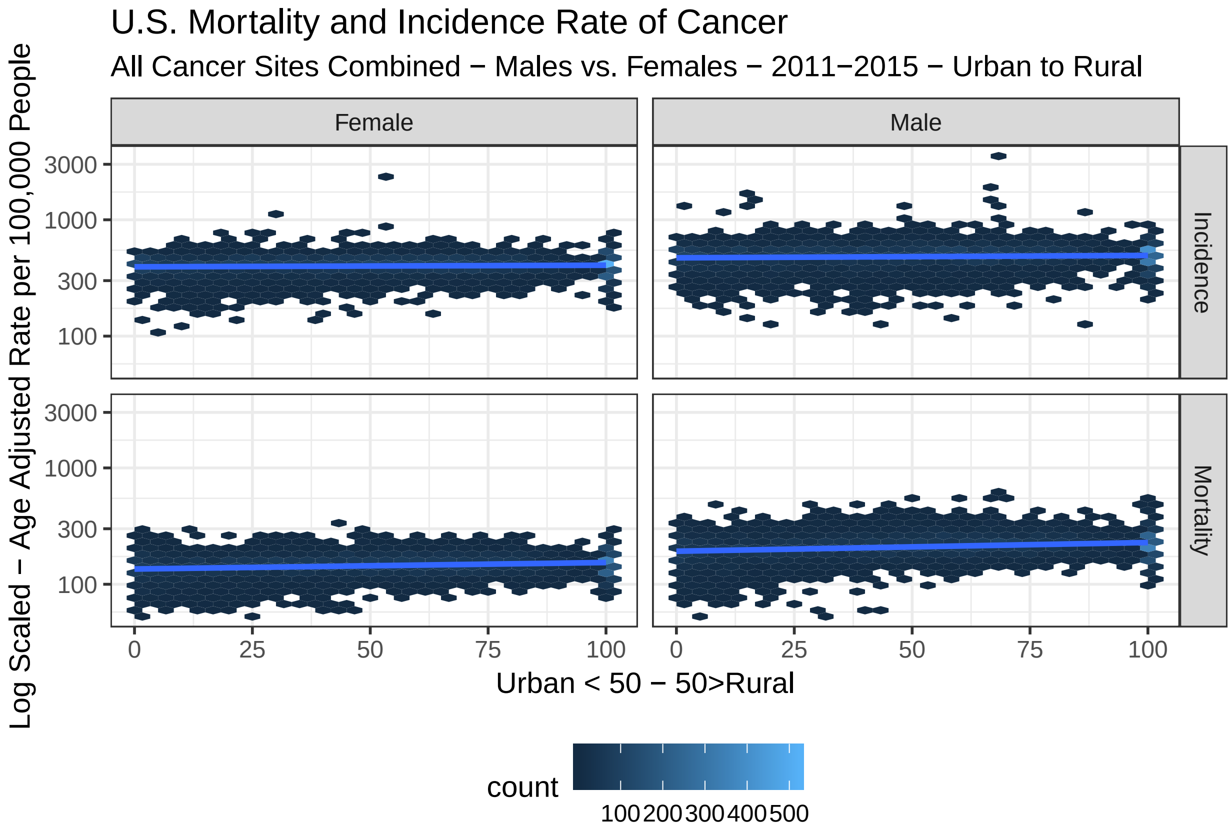
Related Article: 1. Zahnd, W. E., James, A. S., Jenkins, W. D., Izadi, S. R., Fogleman, A. J., Steward, D. E., . Brard, L. (2018). Rural-Urban differences in cancer incidence and trends in the United States. Cancer Epidemiology Biomarkers and Prevention. http://doi.org/10.1158/1055-9965.EPI-17-0430

Summary: *The article describes that although the combined incidence rates were higher in urban areas, their decline was also greater than the rural populations. Most of the discrepancy were related to tobacco-associated, HPV-associated, lung and bronchus, cervical, and colorectal cancers across the population groups.*

By using the byareacounty dataframe which has different area codes for each county, we import another data set from the 2015 GEOID U.S. Census that has the percentage of rural and urban for each county. We join these two dataframes together by the areacode and run a trend analysis using geom_smooth. Although the geom_smooth is highly variable by the variation in the data points, it still shows a trend over the 0-100 Urban to Rural Counties.

```r
byareaCounty %>%
  mutate(areacode = str_extract(AREA, pattern = "\\d+")) %>%
  full_join(rural, by = c("areacode" = "2015 GEOID")) ->
  byareaRural
```

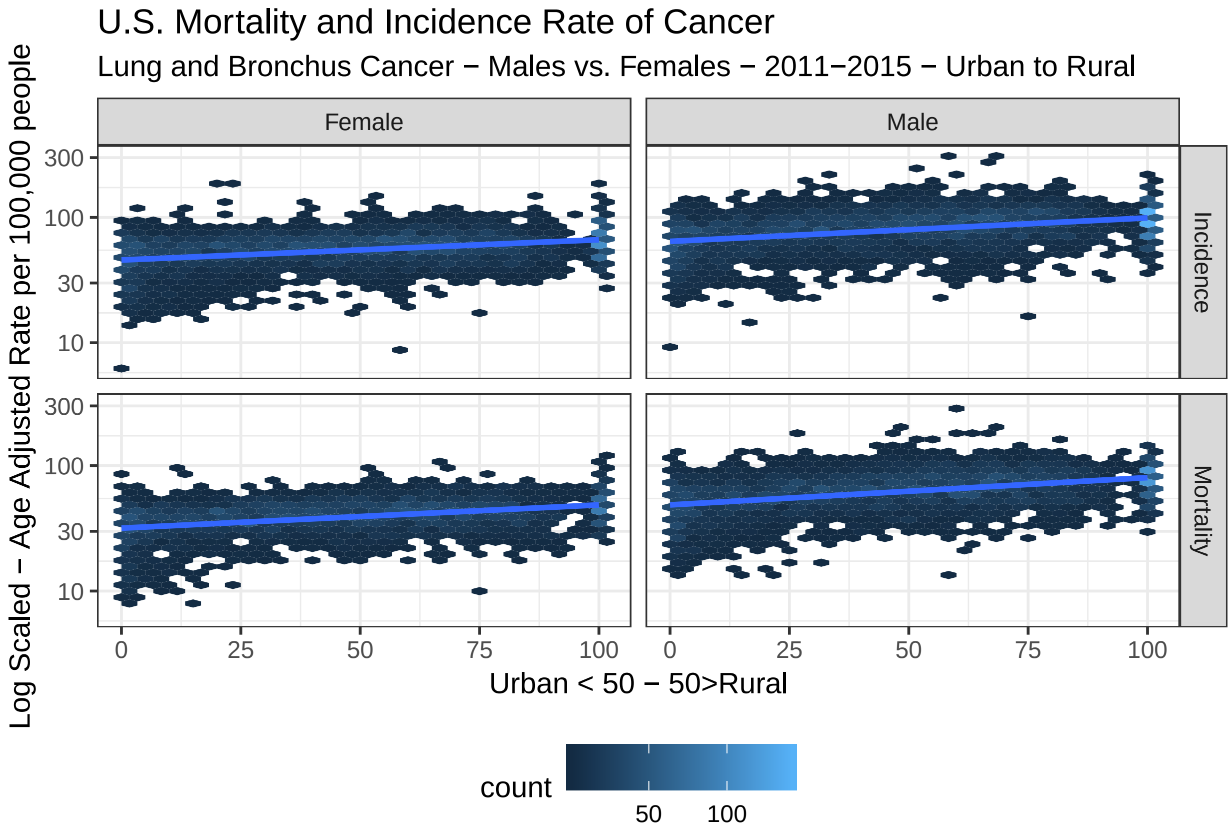
```r
byareaRural%>%
  select(STATE, AREA, areacode, percent, AGE_ADJUSTED_RATE, SITE, SEX, RACE, EVENT_TYPE,
         YEAR)%>%
  filter(!is.na(AGE_ADJUSTED_RATE), !is.na(percent) ,SITE == "All Cancer Sites Combined",
         SEX != "Male and Female") %>%
  ggplot(aes(x = percent, y = AGE_ADJUSTED_RATE)) +
  geom_hex() +
  facet_grid(EVENT_TYPE ~ SEX) +
  scale_y_log10() +
  geom_smooth(method = lm, se = FALSE)+
  theme_bw()+
  labs(title = "U.S. Mortality and Incidence Rate of Cancer",
    y = "Log Scaled - Age Adjusted Rate per 100,000 People",
    x = "Urban < 50 - 50>Rural",
    subtitle = "All Cancer Sites Combined - Males vs. Females - 2011-2015 - Urban to Rural")+
  theme(legend.position = 'bottom')
```



There seems to be a slight increase as the percentage goes higher.

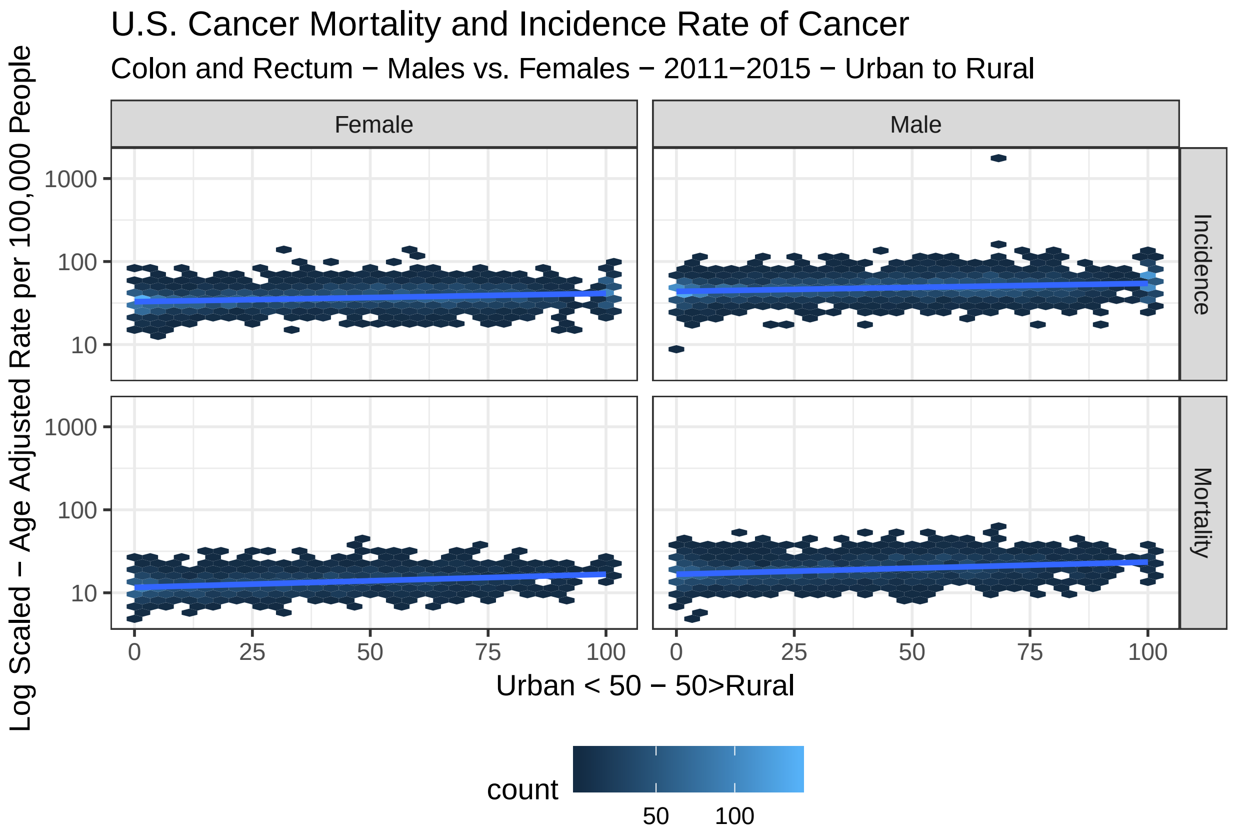
We further analyze this by measuring the rate in the lung and bronchus cancer and also the colon and rectum cancer. We choose mainly these two cancer sites since we're further hypothesizing that rural populations might have high percentage of smokers and thus higher rate of Lung and Bronchus cancer.

```r
byareaRural%>%
  filter(!is.na(AGE_ADJUSTED_RATE),!is.na(percent) , SITE == "Lung and Bronchus",
         SEX != "Male and Female") %>%
  ggplot(aes(x = percent, y = AGE_ADJUSTED_RATE)) +
  geom_hex() +
  facet_grid(EVENT_TYPE ~ SEX) +
  scale_y_log10() +
  geom_smooth(method = lm, se = FALSE)+
  theme_bw()+
    labs(title = "U.S. Mortality and Incidence Rate of Cancer",
    y = "Log Scaled - Age Adjusted Rate per 100,000 people",
    x = "Urban < 50 - 50>Rural",
    subtitle = "Lung and Bronchus Cancer - Males vs. Females - 2011-2015 - Urban to Rural")+
  theme(legend.position = 'bottom')
```



As seen by this plot, there's an upward trend as the rural percentage goes higher.

```
byareaRural%>%
  filter(SITE == "Colon and Rectum", !is.na(AGE_ADJUSTED_RATE),!is.na(percent),
         SEX != "Male and Female") %>%
  ggplot(aes(x = percent, y = AGE_ADJUSTED_RATE)) +
  geom_hex() +
  facet_grid(EVENT_TYPE ~ SEX) +
  scale_y_log10() +
  geom_smooth(method = lm, se = FALSE)+
  theme_bw()+
    labs(title = "U.S. Cancer Mortality and Incidence Rate of Cancer",
    y = "Log Scaled - Age Adjusted Rate per 100,000 People",
    x = "Urban < 50 - 50>Rural",
    subtitle = "Colon and Rectum - Males vs. Females - 2011-2015 - Urban to Rural")+
  theme(legend.position = 'bottom')
```



We can also interpret that for the colon and rectum cancer there's an slightly upward trend as the rural percentage goes higher, yet slope is less than the lung and bronchus.

HYPOTHESIS 2

The mortality rate of cancer in elderlies are higher than the other age groups.

Related Article: 2. White, M. C., Holman, D. M., Boehm, J. E., Peipins, L. A., Grossman, M., & Jane Henley, S. (2014). Age and cancer risk: A potentially modifiable relationship. American Journal of Preventive Medicine. http://doi.org/10.1016/j.amepre.2013.10.029

Summary: *After midlife the frequency of several cancer risk factors and the incidence rate begin to increase.*

Using the byage dataframe, we create two groups of elderlies and non-elderlies and by filtering out for mortality in all cancer sites combined, we visualize a boxplot showing the difference in the rate.

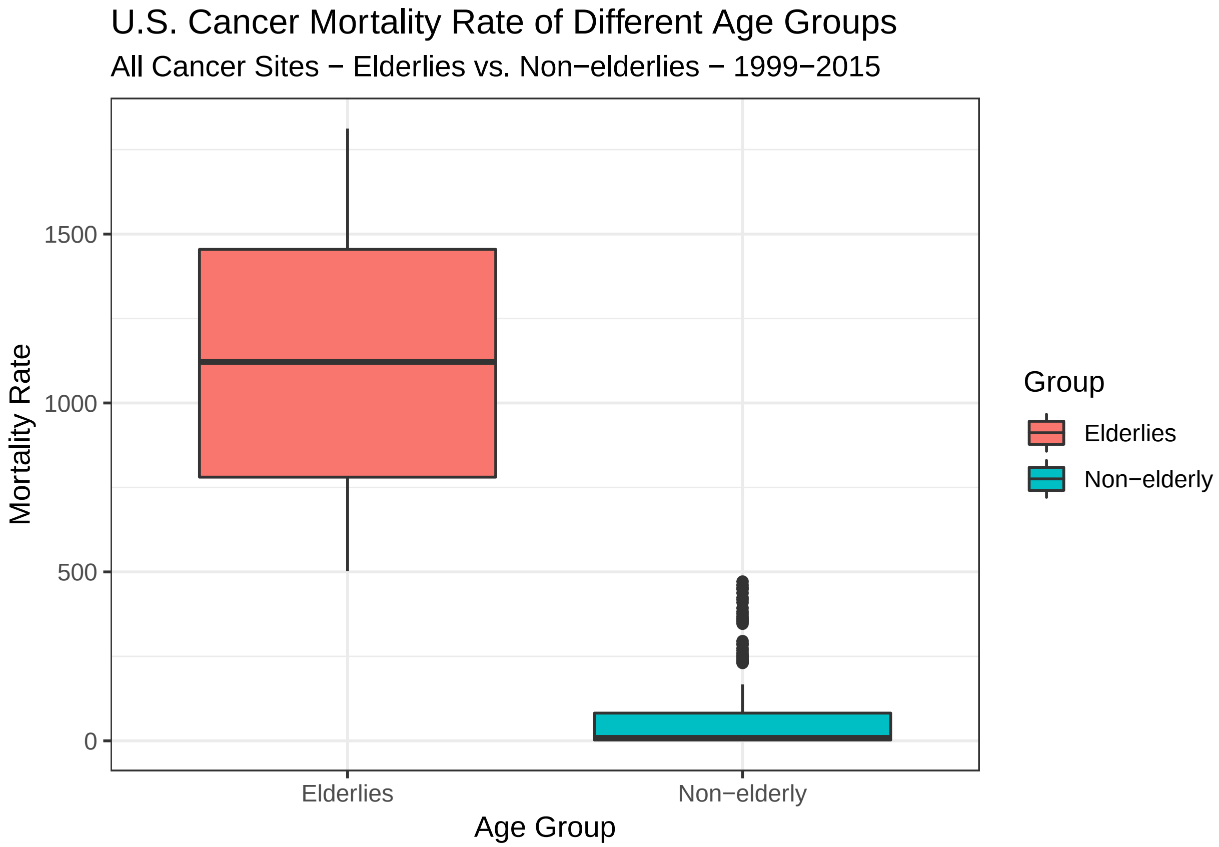
```
byage %>%#Taking out the 2011-2015 Grouping which is duplicate and not useful
  filter(is.na(YEAR)==FALSE)->
  byage

#Creating a group for the elderlies
byage %>%
  filter(AGE == "65-69" | AGE == "70-74" | AGE == "75-79" | AGE ==  "80-84" |
           AGE ==  "85+") %>%
  mutate(Group = "Elderlies")->
  byage1

#Creating a group for the rest or non-elderly
byage%>%
  filter(AGE == "<1" |AGE== "1-4" |AGE== "5-9" |AGE== "10-14" |AGE== "15-19" |
           AGE== "20-24" |AGE== "25-29" |AGE== "30-34" |AGE== "35-39" |
           AGE== "40-44" |AGE== "45-49" |AGE== "50-54" |AGE== "55-59" |AGE== "60-64") %>%
  mutate(Group = "Non-elderly")->
  byage2

rbind(byage1, byage2) -> byage3

byage3%>%
  filter(EVENT_TYPE == "Mortality", SEX == "Male and Female", RACE == "All Races",
         !is.na(RATE), SITE == "All Cancer Sites Combined")%>%
  ggplot(mapping = aes(Group, RATE, fill = Group))+
  geom_boxplot()+
  theme_bw()+
  labs(title = "U.S. Cancer Mortality Rate of Different Age Groups",
    y = "Mortality Rate",
    x = "Age Group",
    subtitle = "All Cancer Sites - Elderlies vs. Non-elderlies - 1999-2015")
```



After getting the boxplot which clearly shows the difference between the mortality of elderlies and the rest, we try to measure the `Death per Incidence` rate of each group. We define this value as `Mortality/Incidence*100`. Although this value is not really usable in terms of the same deaths being related to the same incidences, it can still help us in determining a better proportion of death per each year.

```
#Spreading the count value for incidence and mortality
#in order to have them in separate columns for measuring (Incidence - Mortality) later.
byage3 %>%
  spread(key = EVENT_TYPE, value = COUNT)->
  byage3s

#arranging the values to put the same year, age and group next to each other
byage3s%>%
  select(AGE, Group, YEAR, Mortality, Incidence, SEX, RACE, SITE)%>%
    arrange(YEAR, AGE)->
  byage3s

#Taking out NA values
byage3s%>%
  select(-Incidence)%>%
  filter(!is.na(Mortality)) -> byage4

#Taking out NA values
byage3s%>%
  select(-Mortality)%>%
  filter(!is.na(Incidence)) -> byage5
```

```r
#Joining the dataframes for visualization
full_join(byage4, byage5) ->byage6
```

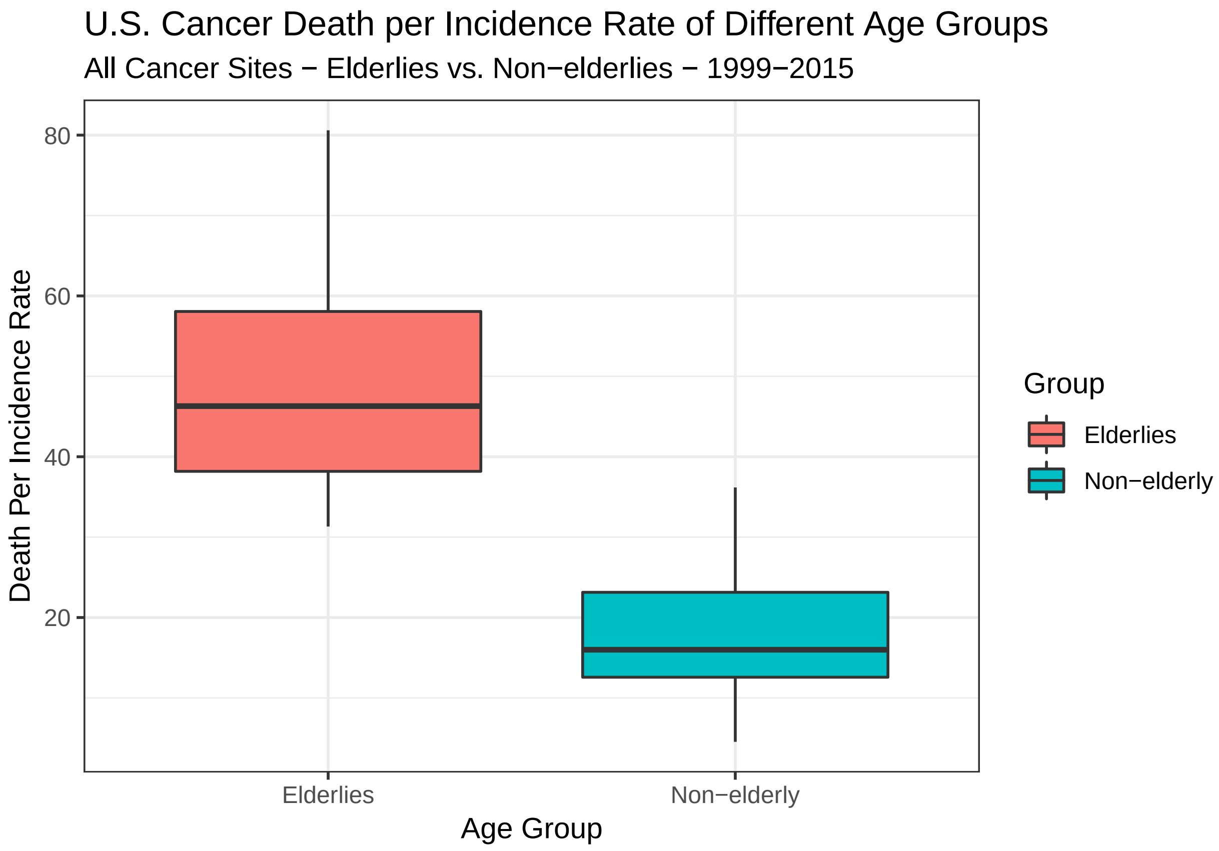
```
## Joining, by = c("AGE", "Group", "YEAR", "SEX", "RACE", "SITE")
```

```r
#Creating the death per incidence rate and parsing the factors for the AGE
byage6%>%
  mutate(dpi = (Mortality/Incidence*100))%>%
  mutate(AGE = parse_factor(AGE)) ->
  byage6
#Ordering the factor levels of AGE
byage6%>%
  mutate(AGE = fct_relevel(AGE, "5-9", after = 2))->
byage6

#checking the order
levels(byage6$AGE)
```

```
##  [1] "<1"    "1-4"   "5-9"   "10-14" "15-19" "20-24" "25-29" "30-34"
##  [9] "35-39" "40-44" "45-49" "50-54" "55-59" "60-64" "65-69" "70-74"
## [17] "75-79" "80-84" "85+"
```

```r
#plotting
byage6%>%
  filter(SEX == "Male and Female", RACE == "All Races", SITE == "All Cancer Sites Combined",
         !is.na(dpi))%>%
    ggplot(mapping = aes(Group, dpi, fill = Group))+
  geom_boxplot()+
  theme_bw()+
    labs(title = "U.S. Cancer Death per Incidence Rate of Different Age Groups",
    y = "Death Per Incidence Rate",
    x = "Age Group",
    subtitle = "All Cancer Sites - Elderlies vs. Non-elderlies - 1999-2015")
```



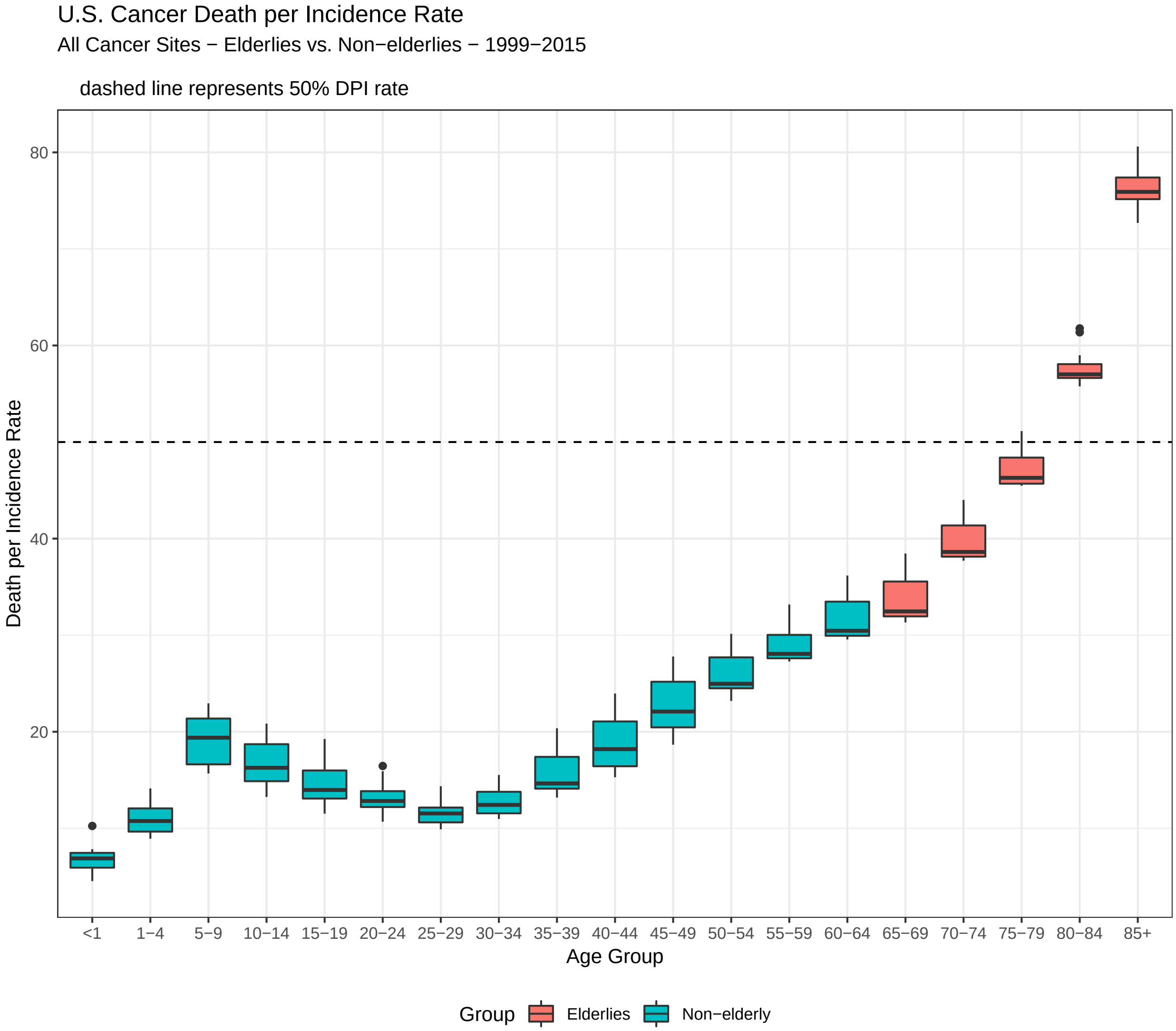
```r
#Printing it out the median
byage6%>%
  filter(SEX == "Male and Female", RACE == "All Races",
         SITE == "All Cancer Sites Combined", !is.na(dpi))%>%
  group_by(Group)%>%
  summarize(median = median(dpi))
```

```
## # A tibble: 2 x 2
##   Group       median
##   <chr>        <dbl>
## 1 Elderlies     46.3
## 2 Non-elderly   16.0
```

As we can see there is a strong association between the age group and the survival rate. While the median for Elderlies are 46.29, same for the rest of the age groups are 15.99.

We try to do a much more detailed analysis for all age groups:

```r
byage6%>%
  filter(SEX == "Male and Female", RACE == "All Races",
         SITE == "All Cancer Sites Combined", !is.na(dpi))%>%
    ggplot(mapping = aes(AGE, dpi, fill = Group))+
  geom_boxplot()+
  theme_bw()+
    labs(title = "U.S. Cancer Death per Incidence Rate",
    y = "Death per Incidence Rate",
    x = "Age Group",
    subtitle = "All Cancer Sites - Elderlies vs. Non-elderlies - 1999-2015 \n
    dashed line represents 50% DPI rate")+
  geom_hline(yintercept = 50, lty = 2, col = "black")+
  theme(legend.position = 'bottom')
```



The resulting plot shows that all elderly groups had a higher dpi rate than the non-elderlies. The horizontal dashed line represents the 50% rate.

An interesting finding is that the dpi rate of the “25-29” group is the lowest between “5-85+”.

Addtionally, age groups of 0-4 have the lowest dpi in general. This might be related to the specific high incidence cancer types in these age groups such as leukemias, brain and central nervous system tumors, and lymphomas which due to the progress in cancer treatments are providing a lower mortality. Yet as the age goes up, the number of cancer types humans can get generally increases.

HYPOTHESIS 3

There's an association between the death rate of skin cancer and different ethnicities.

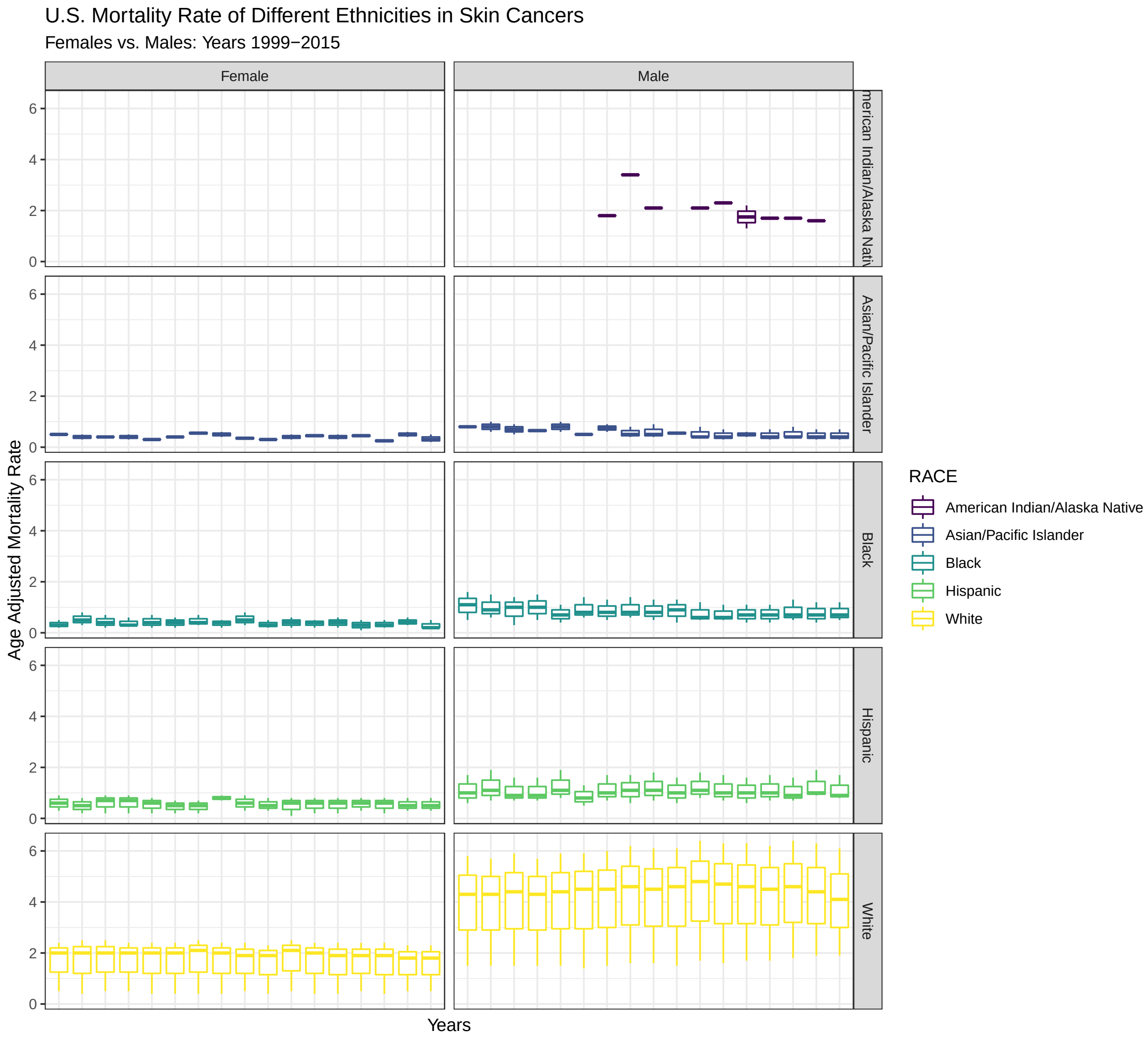
Related Article: 3. Ward-Peterson, M., Acuna, J. M., Alkhalifah, M. K., Nasiri, A. M., Al-Akeel, E. S., Alkhaldi, T. M., Dawari, S. A., . Aldaham, S. A. (2016). Association Between Race/Ethnicity and Survival of Melanoma Patients in the United States Over 3 Decades: A Secondary Analysis of SEER Data. Medicine, 95(17), e3315.

Summary: *The age groups of 18+ were diagnosed with primary cutaneous melanoma from 1982 to 2011. Considering the cause specific mortality and controlling for stage and site, non-Hispanic Black ethnicity had a lower Hazard Rate compared to other populations such as non-Hispanic Whites.*

We have three skin related cancer types in the bysite dataframe:

1. Melanomas of the Skin
2. Other Non-Epithelial Skin
3. Skin excluding Basal and Squamous

```
bysite %>%
  filter(str_detect(SITE, "(?i)skin")) %>%
  filter(EVENT_TYPE == "Mortality", SEX != "Male and Female",
         RACE != "All Races", YEAR != "2011-2015", !is.na(AGE_ADJUSTED_RATE)) %>%
  group_by(RACE)%>%
  ggplot( mapping = aes(y = AGE_ADJUSTED_RATE, x = YEAR, col = RACE))+
  geom_boxplot()+
  facet_grid(RACE~SEX)+
  theme_bw()+
  labs(title = "U.S. Mortality Rate of Different Ethnicities in Skin Cancers",
    y = "Age Adjusted Mortality Rate",
    x = "Years",
    subtitle = "Females vs. Males: Years 1999-2015")+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())+
  scale_colour_viridis_d(option  = "viridis")
```



```
bysite%>%
  filter(str_detect(SITE, "(?i)skin"), EVENT_TYPE == "Mortality",
         SEX == "Female",RACE == "American Indian/Alaska Native",
         YEAR != "2011-2015", !is.na(AGE_ADJUSTED_RATE))
```

```
## # A tibble: 0 x 13
## # ... with 13 variables: YEAR <chr>, RACE <chr>, SEX <chr>, SITE <chr>,
## #   EVENT_TYPE <chr>, AGE_ADJUSTED_CI_LOWER <chr>,
## #   AGE_ADJUSTED_CI_UPPER <chr>, AGE_ADJUSTED_RATE <dbl>, COUNT <chr>,
## #   POPULATION <dbl>, CRUDE_CI_LOWER <chr>, CRUDE_CI_UPPER <chr>,
## #   CRUDE_RATE <chr>
```

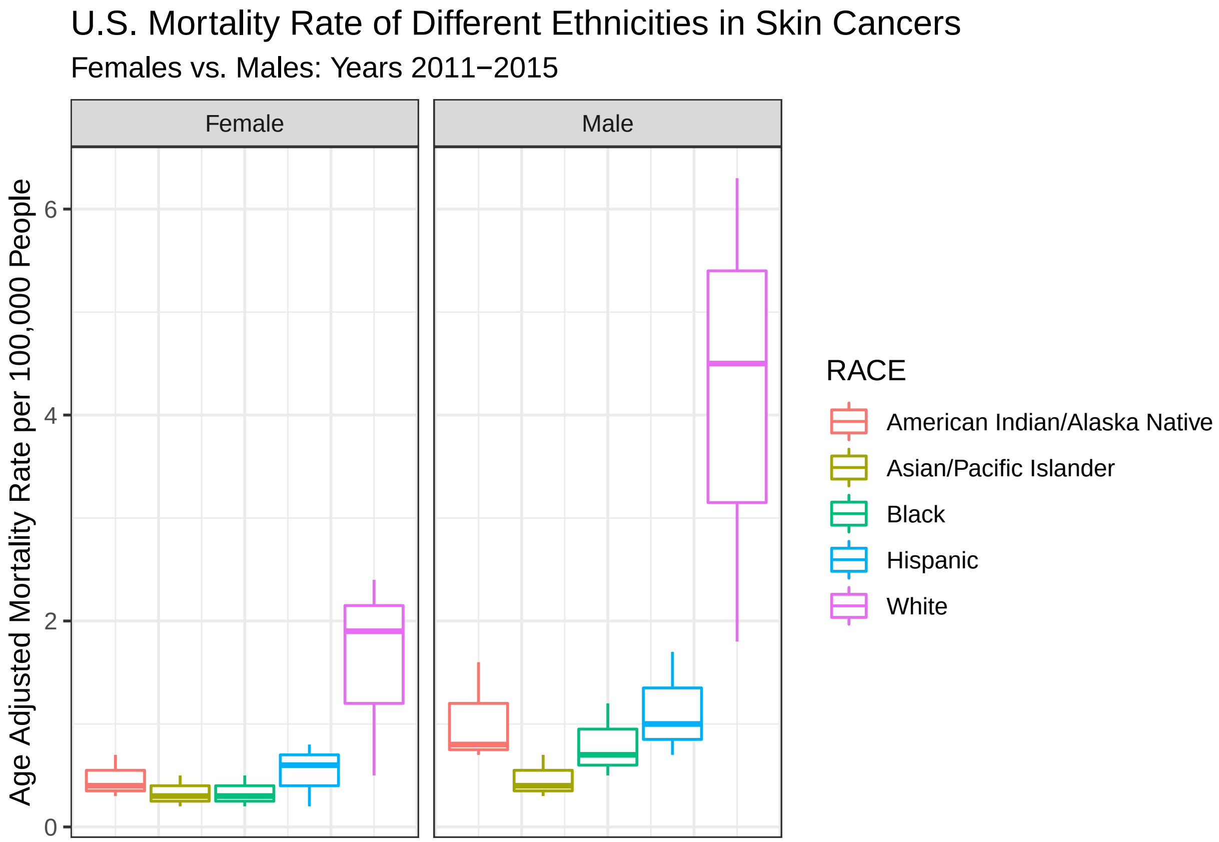
There's no mortality rate recorded for "American Indian/Alaska Native" Females per each year from 1999-2015, yet we can see there are few recordings for the specific years of 2011-2015.

```r
bysite%>%
  filter(str_detect(SITE, "(?i)skin"), EVENT_TYPE == "Mortality",
         SEX == "Female",RACE == "American Indian/Alaska Native",
         YEAR == "2011-2015", !is.na(AGE_ADJUSTED_RATE))
```

```
## # A tibble: 3 x 13
##   YEAR  RACE  SEX   SITE  EVENT_TYPE AGE_ADJUSTED_CI~ AGE_ADJUSTED_CI~
##   <chr> <chr> <chr> <chr> <chr>      <chr>            <chr>
## 1 2011~ Amer~ Fema~ Mela~ Mortality  0.3              0.6
## 2 2011~ Amer~ Fema~ Othe~ Mortality  0.2              0.4
## 3 2011~ Amer~ Fema~ Skin~ Mortality  0.5              0.9
## # ... with 6 more variables: AGE_ADJUSTED_RATE <dbl>, COUNT <chr>,
## #   POPULATION <dbl>, CRUDE_CI_LOWER <chr>, CRUDE_CI_UPPER <chr>,
## #   CRUDE_RATE <chr>
```

Since the the trend in years 1999 - 2015 is similar, for better visualization purposes and to have some values for the American Indian/Alaska Native we limit our visualization to the `2011-2015` data.

```
bysite %>%
  filter(str_detect(SITE, "(?i)skin")) %>%
  filter(EVENT_TYPE == "Mortality", SEX != "Male and Female",
         RACE != "All Races", YEAR == "2011-2015", !is.na(AGE_ADJUSTED_RATE)) %>%
  group_by(RACE)%>%
  ggplot( mapping = aes(y = AGE_ADJUSTED_RATE, col = RACE))+
  geom_boxplot()+
  facet_wrap(vars(SEX))+
  theme_bw()+
  labs(title = "U.S. Mortality Rate of Different Ethnicities in Skin Cancers",
    y = "Age Adjusted Mortality Rate per 100,000 People",
    subtitle = "Females vs. Males: Years 2011-2015")+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



Acording to the resulting plot, there seems to be a much higher skin cancer mortality rate in `White` ethinicity compared to the others. Additionally, males seems to be have a higher mortality rate.

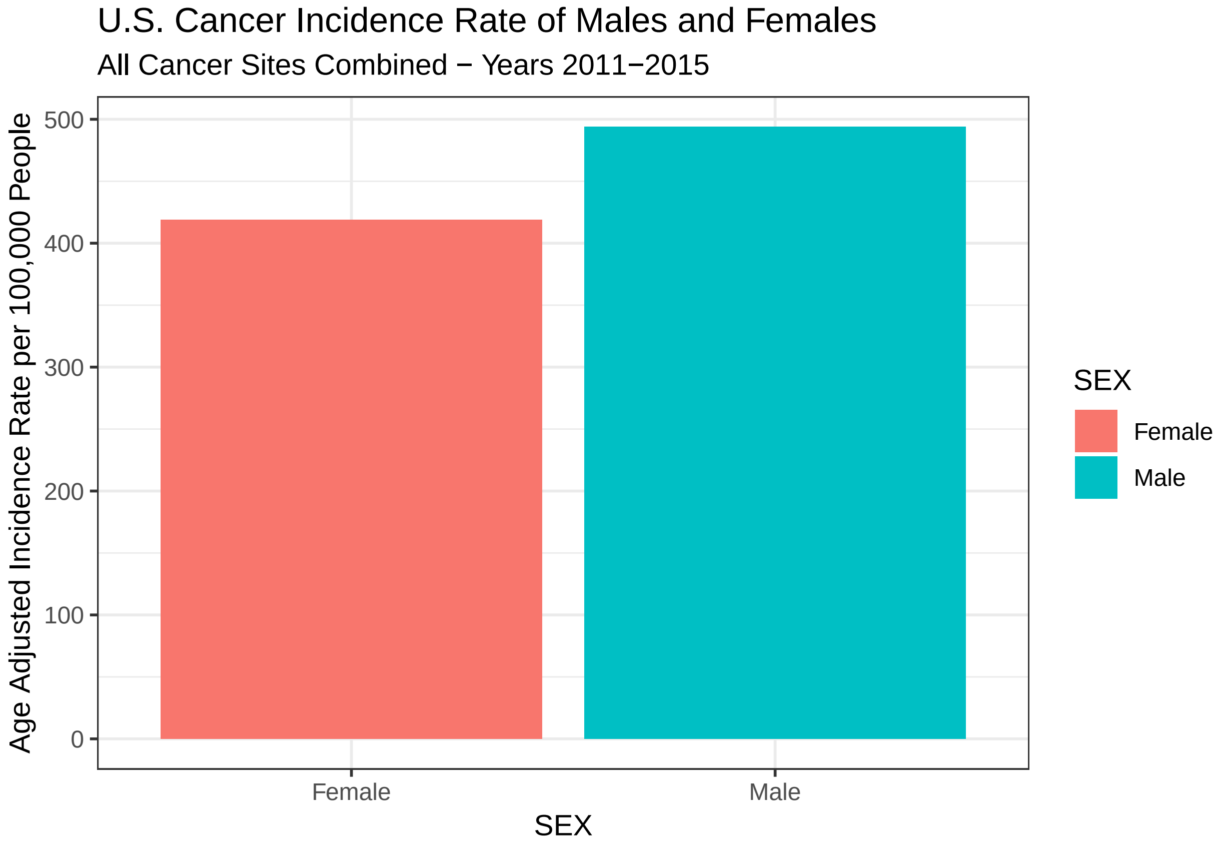
HYPOTHESIS 4

Males are more prone to new cancers than females:

Related Article: 4. Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: A Cancer Journal for Clinicians. http://doi.org/10.3322/caac.21551

Summary: *According to the article, all sites combined in 2011-2015, the incidence rate of Males are 494.8 compared to the 419.3 of the Females. Units are per 100,000 population and age adjusted to the 2000 US standard population.*

```
bysite%>%
  filter(EVENT_TYPE == "Incidence", SITE == "All Cancer Sites Combined",
         YEAR == "2011-2015", RACE == "All Races",
         SEX!= "Male and Female", !is.na(AGE_ADJUSTED_RATE)) %>%
  ggplot(mapping = aes(SEX, AGE_ADJUSTED_RATE, fill = SEX))+
  geom_col()+
  theme_bw()+
  labs(title = "U.S. Cancer Incidence Rate of Males and Females",
    y = "Age Adjusted Incidence Rate per 100,000 People",
    subtitle = "All Cancer Sites Combined - Years 2011-2015")
```



```
bysite%>%
  filter(EVENT_TYPE == "Incidence", SITE == "All Cancer Sites Combined",
         YEAR == "2011-2015", RACE == "All Races", SEX!= "Male and Female", !is.na(AGE_ADJUSTED_RATE))
  group_by(SEX)%>%
  summarize(median = median(AGE_ADJUSTED_RATE))
```

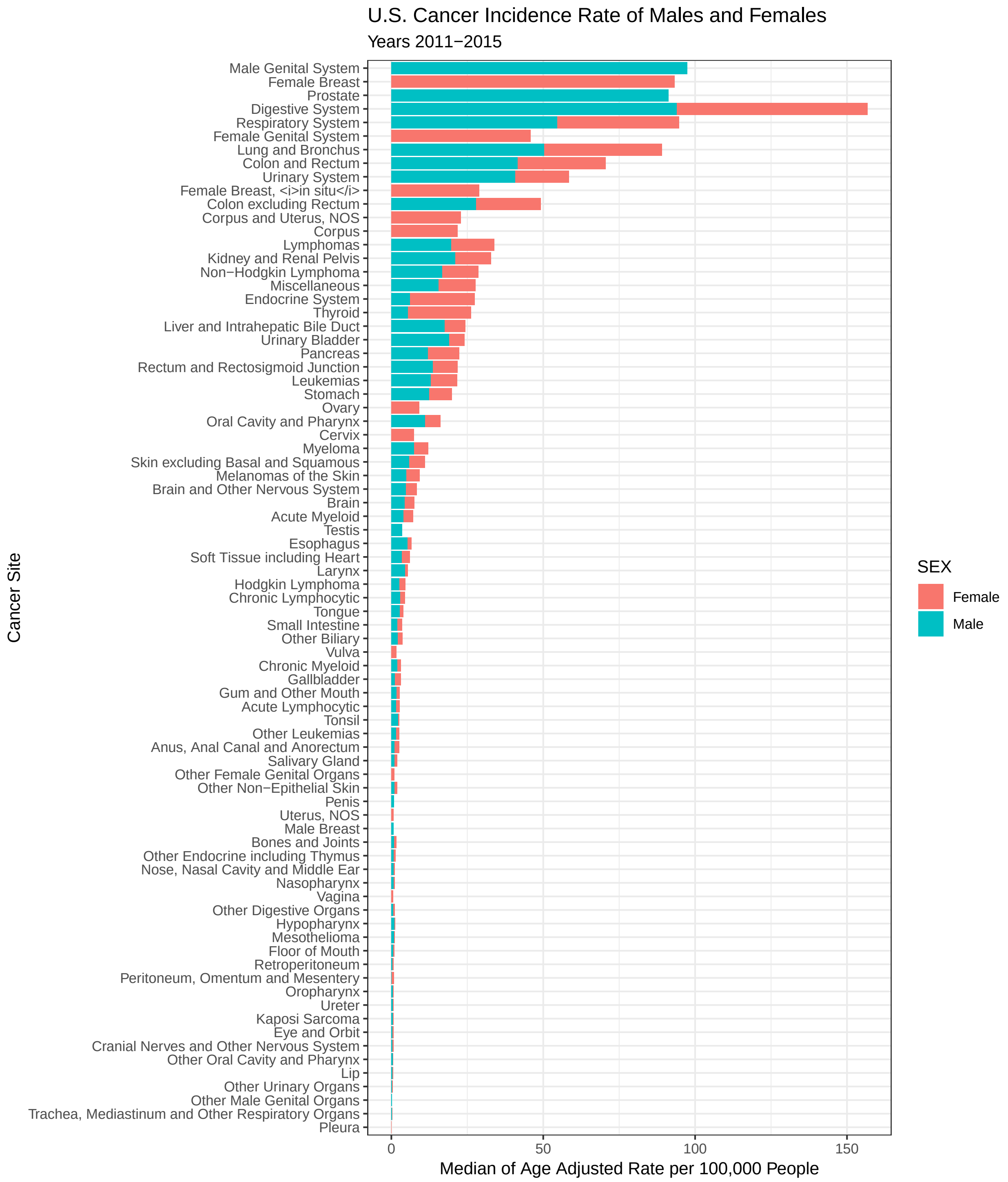
```
## # A tibble: 2 x 2
##   SEX    median
##   <chr>   <dbl>
## 1 Female   419
## 2 Male     494.
```

Resulting plot comes with two values of median of 493.8 - Males and 419 - Females for the incidence rate for all races and all cancer sites combined. This is confirming our hypothesis. Addtionally the numbers we get are close to the findings of our article.

We further try to analyze this difference on different cancer sites and order the sites per the incidence rate.

```
bysite%>%
  filter(EVENT_TYPE == "Incidence", SITE != "All Cancer Sites Combined",
         YEAR == "2011-2015", RACE != "All Races",
         SEX!= "Male and Female",
         !is.na(AGE_ADJUSTED_RATE))%>%
  group_by(SITE, SEX)%>%
  summarize(median = median(AGE_ADJUSTED_RATE)) ->bysitein

bysite%>%
  filter(EVENT_TYPE == "Incidence", SITE != "All Cancer Sites Combined",
         YEAR == "2011-2015", RACE != "All Races", SEX!= "Male and Female", !is.na(AGE_ADJUSTED_RATE))%
  group_by(SITE, SEX)%>%
  summarize(median = median(AGE_ADJUSTED_RATE))%>%
  arrange(desc(median))%>%
  filter(!str_detect(SITE, "All Sites"))%>%
  ggplot(aes(x = reorder(SITE, median), y = median, fill = SEX))+
  geom_col()+
  coord_flip()+
  theme_bw()+
  labs(title = "U.S. Cancer Incidence Rate of Males and Females",
    y = "Median of Age Adjusted Rate per 100,000 People",
    x = "Cancer Site",
    subtitle = "Years 2011-2015")
```

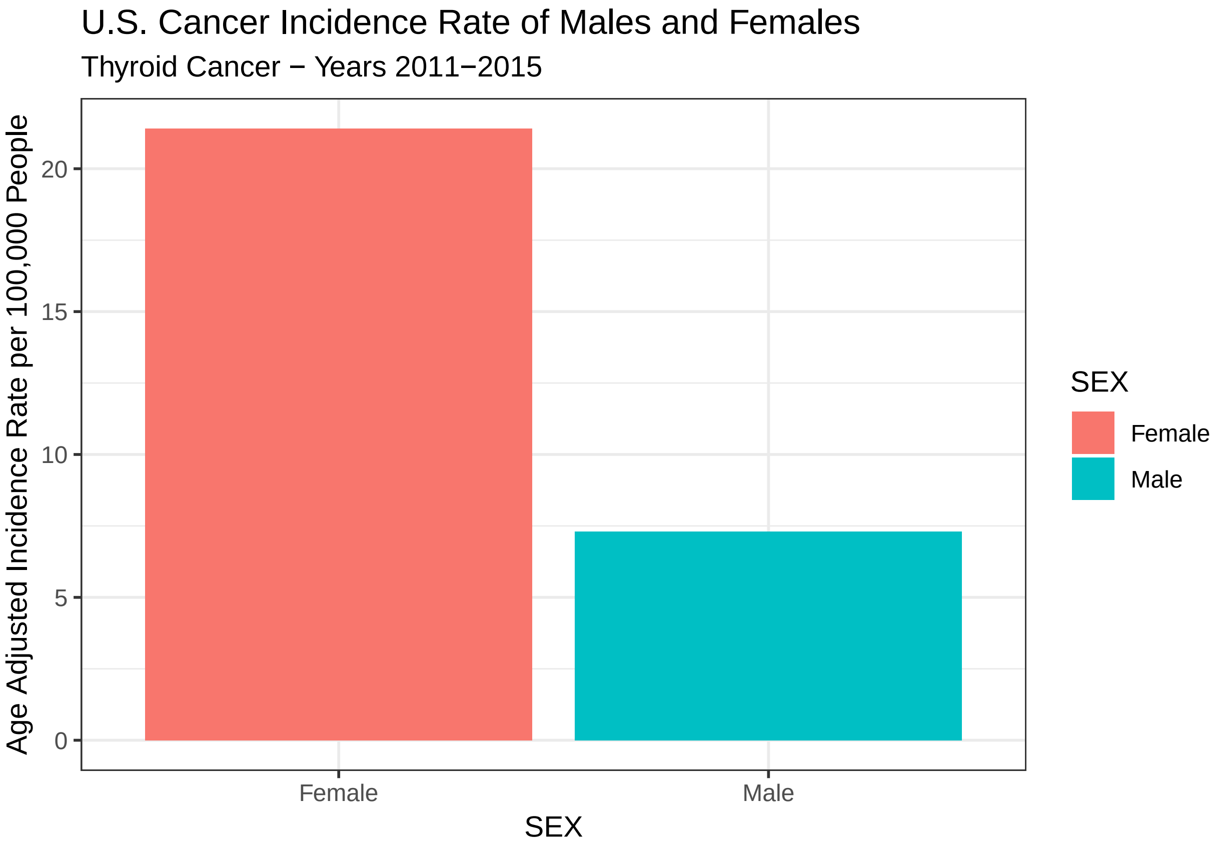


While the resulting plot seems overwhelming, It clearly shows which cancer types have a higher incidence rate for different sexes.

One interesting finding is that with endocrine system and thyroid cancer, female incidence rate is actually higher than males.

Thus we try to analyze this with a better visualization.

```
bysite%>%
  filter(EVENT_TYPE == "Incidence", SITE == "Thyroid", YEAR == "2011-2015",
         RACE == "All Races", SEX!= "Male and Female", !is.na(AGE_ADJUSTED_RATE)) %>%
  ggplot(mapping = aes(SEX, AGE_ADJUSTED_RATE, fill = SEX))+
  geom_col()+
  theme_bw()+
  labs(title = "U.S. Cancer Incidence Rate of Males and Females",
    y = "Age Adjusted Incidence Rate per 100,000 People",
    subtitle = "Thyroid Cancer - Years 2011-2015")
```



```
bysite%>%
  filter(EVENT_TYPE == "Incidence", SITE == "Thyroid", YEAR == "2011-2015",
         RACE == "All Races", SEX!= "Male and Female", !is.na(AGE_ADJUSTED_RATE)) %>%
  group_by(SEX)%>%
  summarize(median = median(AGE_ADJUSTED_RATE))
```

```
## # A tibble: 2 x 2
##   SEX    median
##   <chr>   <dbl>
## 1 Female   21.4
## 2 Male      7.3
```

This is further proven by many research papers done on the gender disparity of thyroid cancer.

Related Article: 5. (Rahbari, R., Zhang, L., & Kebebew, E. (2010). Thyroid cancer gender disparity. Future oncology (London, England), 6(11), 1771-1779. doi:10.2217/fon.10.127)

The paper suggest that this cancer is 2.9 times more common in females. According to our medians of 7.3-F and 21.4-M, we have $21.4/7.3 = 2.93$, which is a good estimate.

HYPOTHESIS 5:

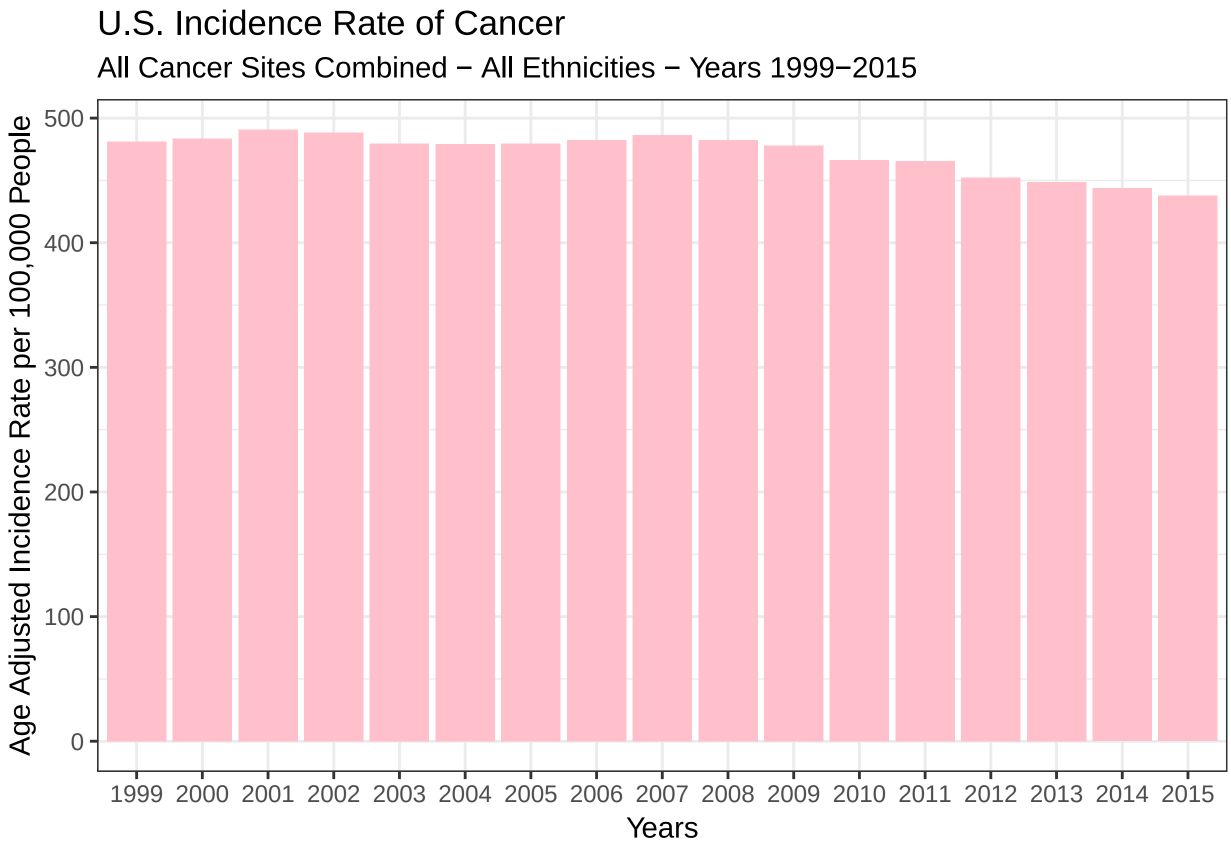
Rate of new cancers during the 1999-2015 should increase.

Related Article: 4. Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: A Cancer Journal for Clinicians. http://doi.org/10.3322/caac.21551

Summary: *Considering all sites combined there is actually a decrease in the incidence rate which can be contributed to the awareness and research done on major lethal cancer groups such as prostate, breast, and lung & bronchus types.*

We visualize the all cancer sites combined incidences between the years 1999-2015.

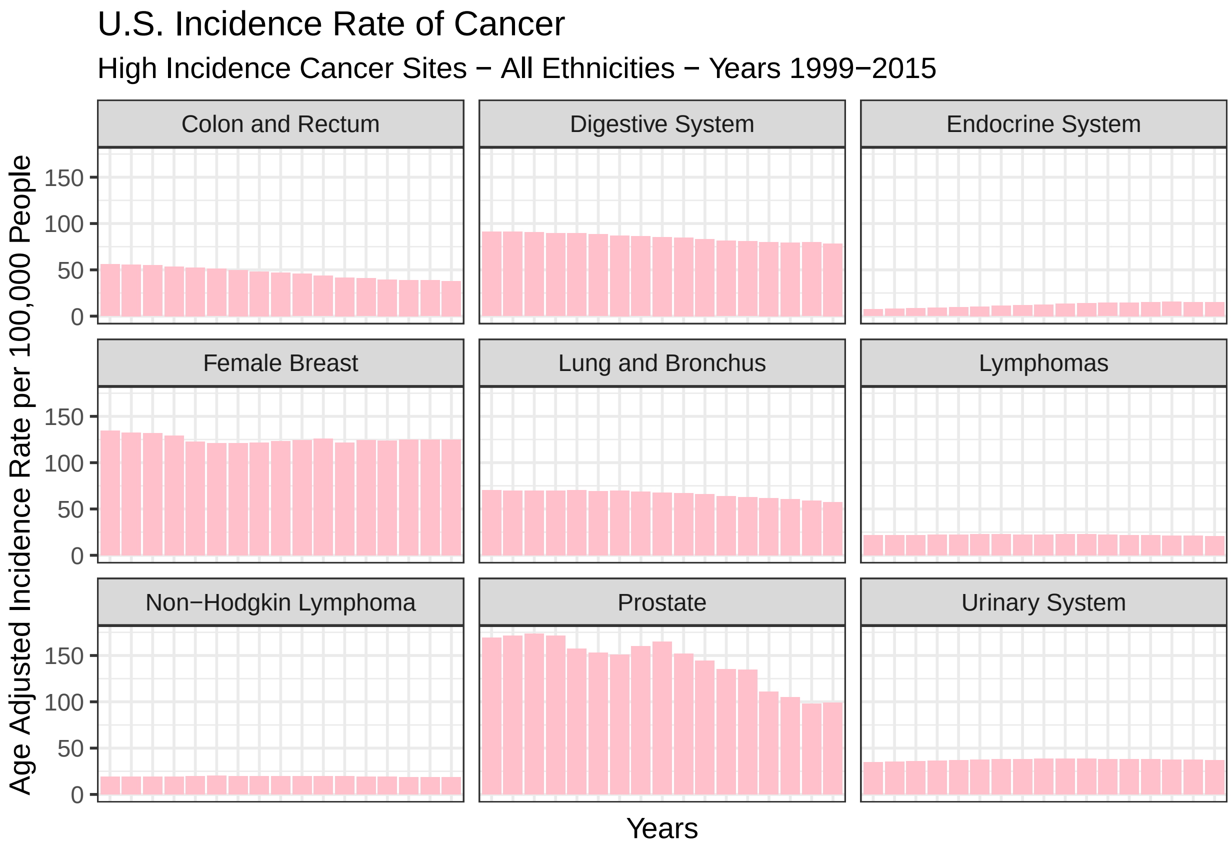
```
bysite %>%
  filter(YEAR != "2011-2015", SITE == "All Cancer Sites Combined", RACE == "All Races",
         SEX == "Male and Female", EVENT_TYPE == "Incidence") %>%
  ggplot(mapping = aes(YEAR, weight = AGE_ADJUSTED_RATE)) +
  geom_bar(fill = "pink")+
  theme_bw()+
  labs(title = "U.S. Incidence Rate of Cancer",
    y = "Age Adjusted Incidence Rate per 100,000 People",
    x = "Years",
    subtitle = "All Cancer Sites Combined - All Ethnicities - Years 1999-2015")
```



Resulting plot nulls our hypothesis since the rate is actually decreasing.

We try to do some analysis on some high incidence cancer sites such as Lung and Bronchus, Female Breast, Prostate, Digestive System, Lymphomas, Urinary Systems, Non-Hodgkin Lymphoma.

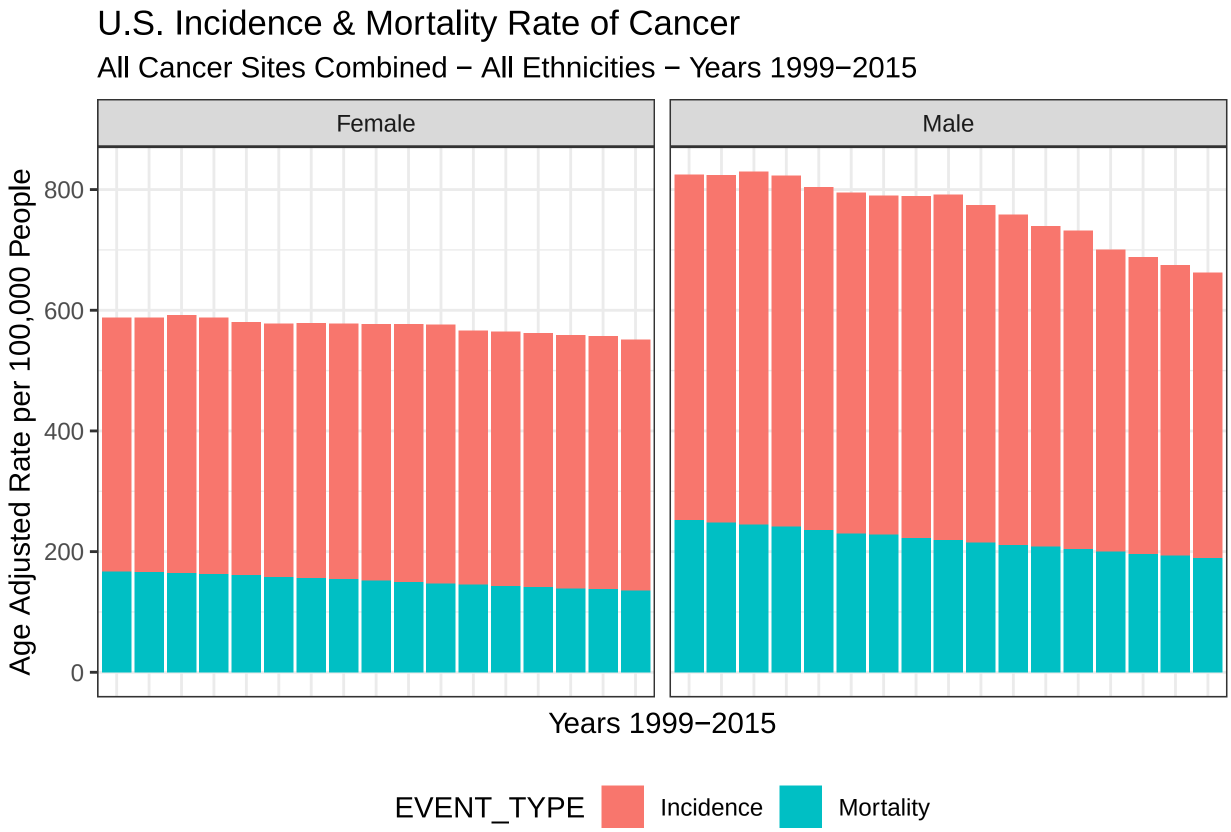
```r
bysite %>%
  filter(YEAR != "2011-2015", SITE == "Lung and Bronchus"|SITE ==  "Female Breast"|
           SITE ==  "Prostate"|SITE ==  "Digestive System"| SITE == "Lymphomas"|
           SITE ==  "Urinary Systems"|SITE ==  "Non-Hodgkin Lymphoma"|
           SITE ==  "Endocrine System"|
           SITE ==  "Colon and Rectum"|SITE ==  "Urinary System",
         RACE == "All Races", SEX == "Male and Female",
         EVENT_TYPE == "Incidence")%>%
  ggplot(mapping = aes(YEAR, weight = AGE_ADJUSTED_RATE)) +
  geom_bar(fill = "pink")+
  theme_bw()+
  facet_wrap(~SITE)+
  labs(title = "U.S. Incidence Rate of Cancer",
    y = "Age Adjusted Incidence Rate per 100,000 People",
    x = "Years",
    subtitle = "High Incidence Cancer Sites - All Ethnicities - Years 1999-2015")+
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



While most sites have a decrease over the years, it seems that the endocrine sytsem cancer had a slight increase over these years.

We further try to analyze both mortality and incidence rate for both sexes separately for all cancer sites combined.

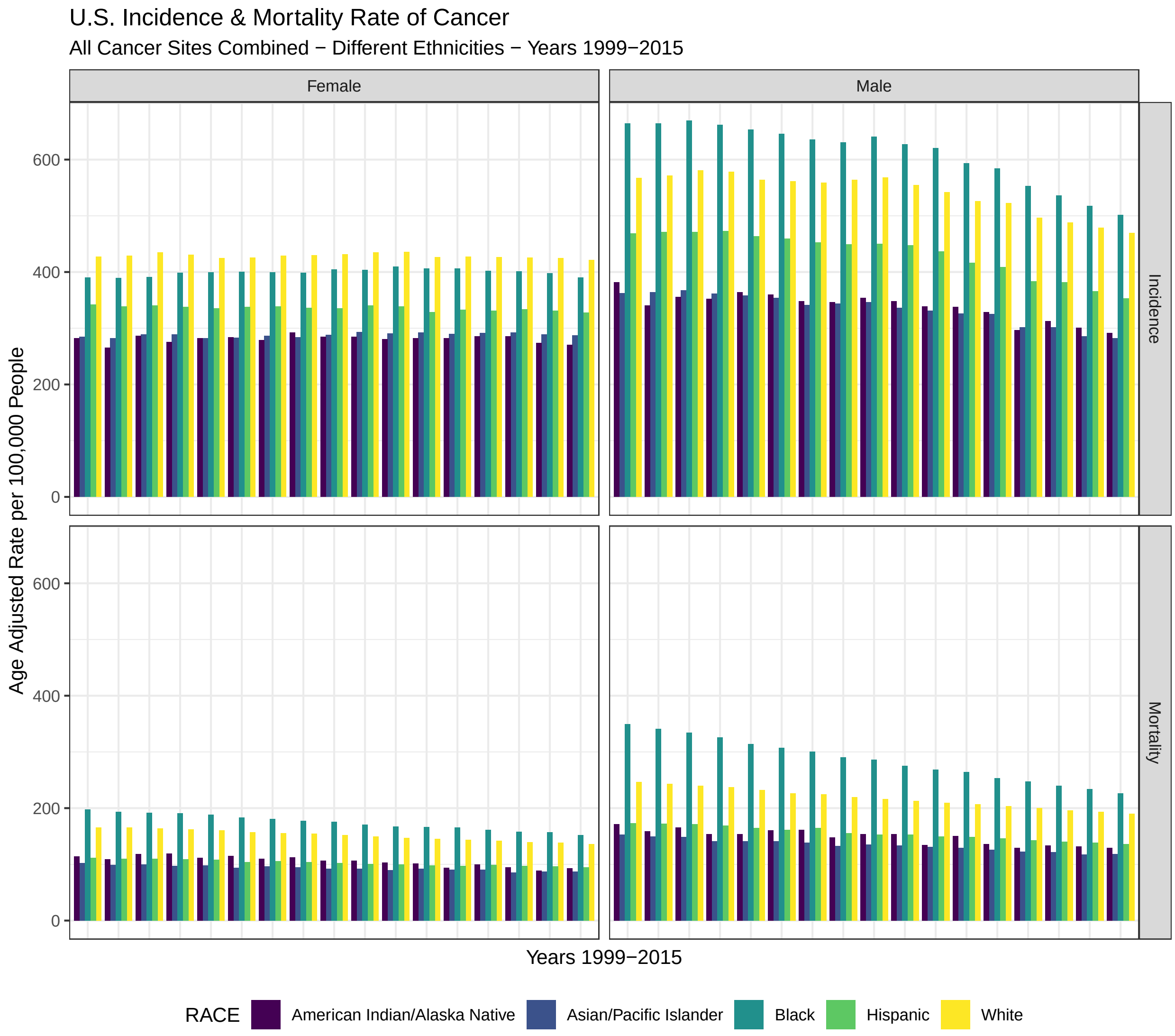
```
bysite%>%
  filter(YEAR != "2011-2015", SEX != "Male and Female",
         SITE == "All Cancer Sites Combined",
         RACE == "All Races")%>%
ggplot(aes(x = YEAR, fill = EVENT_TYPE, weight = AGE_ADJUSTED_RATE))+
  geom_bar() +
  theme_bw() +
  theme(legend.position = 'bottom') +
  facet_wrap(vars(SEX))+
  labs(title = "U.S. Incidence & Mortality Rate of Cancer",
    y = "Age Adjusted Rate per 100,000 People",
    x = "Years 1999-2015",
    subtitle = "All Cancer Sites Combined - All Ethnicities - Years 1999-2015")+
    theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



Resulting plot shows that there's a significant disparity between both the mortality and incidence rate of males vs. females. Again, both trends over the 15 years have decreased.

We then decide to measure this disparity for different ethnicities for all cancer sites.

```
bysite%>%
  filter(YEAR != "2011-2015", SEX != "Male and Female",
         SITE == "All Cancer Sites Combined",
         RACE != "All Races")%>%
  ggplot(aes(x = YEAR, fill = RACE, weight = AGE_ADJUSTED_RATE)) +
  geom_bar(position = "dodge") +
  scale_fill_viridis_d(option  = "viridis") +
  theme_bw() +
  theme(legend.position = 'bottom') +
  facet_grid(EVENT_TYPE~SEX)+
  labs(title = "U.S. Incidence & Mortality Rate of Cancer",
    y = "Age Adjusted Rate per 100,000 People",
    x = "Years 1999-2015",
    subtitle = "All Cancer Sites Combined - Different Ethnicities - Years 1999-2015")+
    theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



One interesting finding is that both the mortality and incidence rate of `Black`s has been higher than other ethnicities.

We further decide to visualize the all cancer sites combined incidence rates for 2011-2015 on different states of the U.S.

```
#importing a state name and abbreviation for joining the state names
abbstate <- read_csv("../FP/STATES.csv")
```

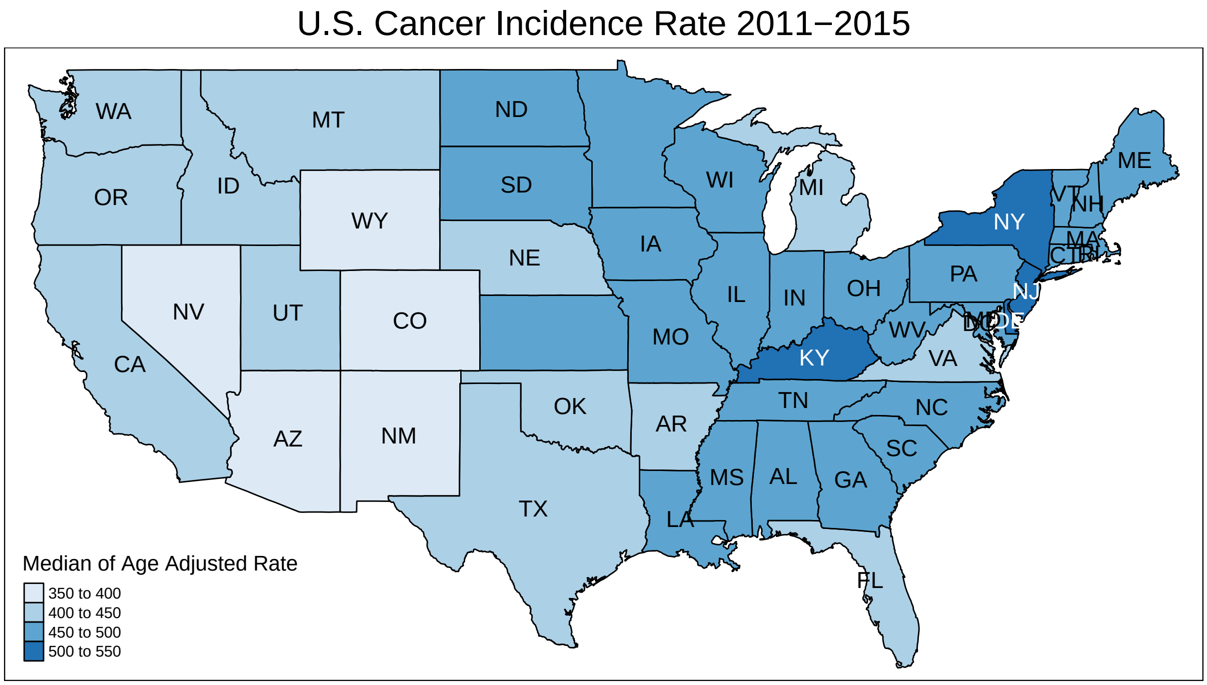
```
## Parsed with column specification:
## cols(
##   name = col_character(),
##   abb = col_character()
## )
```

```
#doing lowercase on state names for matching them later with statedf$ID
abbstate%>%
  mutate(name = sapply(name, tolower)) ->
  abbstate

byareaRural %>%
  full_join(abbstate, by = c("STATE" = "abb")) ->
  byareastate
statedf <- st_as_sf((map("state", plot = FALSE, fill = TRUE)))

byareastate%>%
  filter(SEX == "Male and Female", SITE == "All Cancer Sites Combined",
         RACE == "All Races", EVENT_TYPE == "Incidence")%>%
  group_by(name, State)%>%
  summarize(median = median(AGE_ADJUSTED_RATE, na.rm = TRUE))->
  byareastatesum
```

```r
statedf%>%
  full_join(byareastatesum, by = c("ID" = "name"))%>%
  filter(!is.na(median))%>%
    tm_shape() +
  tm_borders()+
  tm_fill("median", title = "Median of Age Adjusted Rate")+
  tm_text("State")+
  tm_style("col_blind")+
  tm_layout(main.title = "U.S. Cancer Incidence Rate 2011-2015",
            main.title.position = "center")
```



We can clearly see the age adjusted rate for states such as `KY`, `NY`, `NJ`, and `DE` are among the highest for all sites combined. Further research might find some correlations between the conditions of diet, lifestyle, and other variations of the different states.

Conclusion & Discussion

Based on our findings in the relativity of rural and areas and their cancer incidence rate, further research on specific cancer sites might provide useful data to find correlations and apply proper preventive measures. While there was a slight increase in all cancer sites combined towards the rural areas, specific cancer sites such as lung and bronchus, had a much more upward trend. Such data is helpful when analyzing mortality or incidence rates of a specific area and looking for associations such as the diet, life style, and smoking and drinking habits.

Mortality rate association in different age groups can assist in applying preventive measures for the patients and also provide them with a more accurate survival rate based on their cancer type. This can also point us to areas that require better patient care and more extensive research for the cause of death withing specific age groups.

Same can be said for different ethinicities and sexes and their relationship with different cancer type mortality & incidence rates. One can greatly analyze this causal relationship per each cancer site and come up with different therapy solutions since there might be a correlation on whether a specific ethnicity or sex responds to a specific therapy. As we visualized there was a significant disparity in the mortality and incidence rates of males vs. females. It was also interesting that in all cancer sites combined, how the mortality rate of black ethnicity whether male or female was always higher than the other races, but while the incidence of black males were higher, the incidence of black females was actually lower than the whites. .

Analyzing the trends of cancer mortality and incidence rate over the last year, can greatly benefit us in how we approach to cure or prevent each cancer site and whether there's room for improvement for the applied techniques over the last 15 years.. We could visualize that over the past few years, all cancer sites combined had a slight decrease in both incidence and mortality yet, there hasn't been much of a change in the incidence rate of female breast cancer. However, the prostate cancer had a decreae in the incidence rate.

Although cancer has many different sites and is related to many different factors such as life style, sex, ethinicty, living area, and etc., with proper data analysis we can further find correlations in how to deal with each cancer type.

References

1. Zahnd, W. E., James, A. S., Jenkins, W. D., Izadi, S. R., Fogleman, A. J., Steward, D. E., . Brard, L. (2018). Rural-Urban differences in cancer incidence and trends in the United States. Cancer Epidemiology Biomarkers and Prevention. http://doi.org/10.1158/1055-9965.EPI-17-0430

2. White, M. C., Holman, D. M., Boehm, J. E., Peipins, L. A., Grossman, M., & Jane Henley, S. (2014). Age and cancer risk: A potentially modifiable relationship. American Journal of Preventive Medicine. http://doi.org/10.1016/j.amepre.2013.10.029

3. Ward-Peterson, M., Acuna, J. M., Alkhalifah, M. K., Nasiri, A. M., Al-Akeel, E. S., Alkhaldi, T. M., Dawari, S. A., . Aldaham, S. A. (2016). Association Between Race/Ethnicity and Survival of Melanoma Patients in the United States Over 3 Decades: A Secondary Analysis of SEER Data. Medicine, 95(17), e3315.

4. Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. CA: A Cancer Journal for Clinicians. http://doi.org/10.3322/caac.21551

5. (Rahbari, R., Zhang, L., & Kebebew, E. (2010). Thyroid cancer gender disparity. Future oncology (London, England), 6(11), 1771-1779. doi:10.2217/fon.10.127)