DATA 608 Module 3

Jun Pan

3/11/2019

Overview Data

```{r}
cdc <-read.csv("https://raw.githubusercontent.com/johnpannyc/Jun-Pan-DATA-608-Project-3/master/cleaned-cdc-mortality-1999-2010-2.csv")
```

```{r}
glimpse(cdc)
```

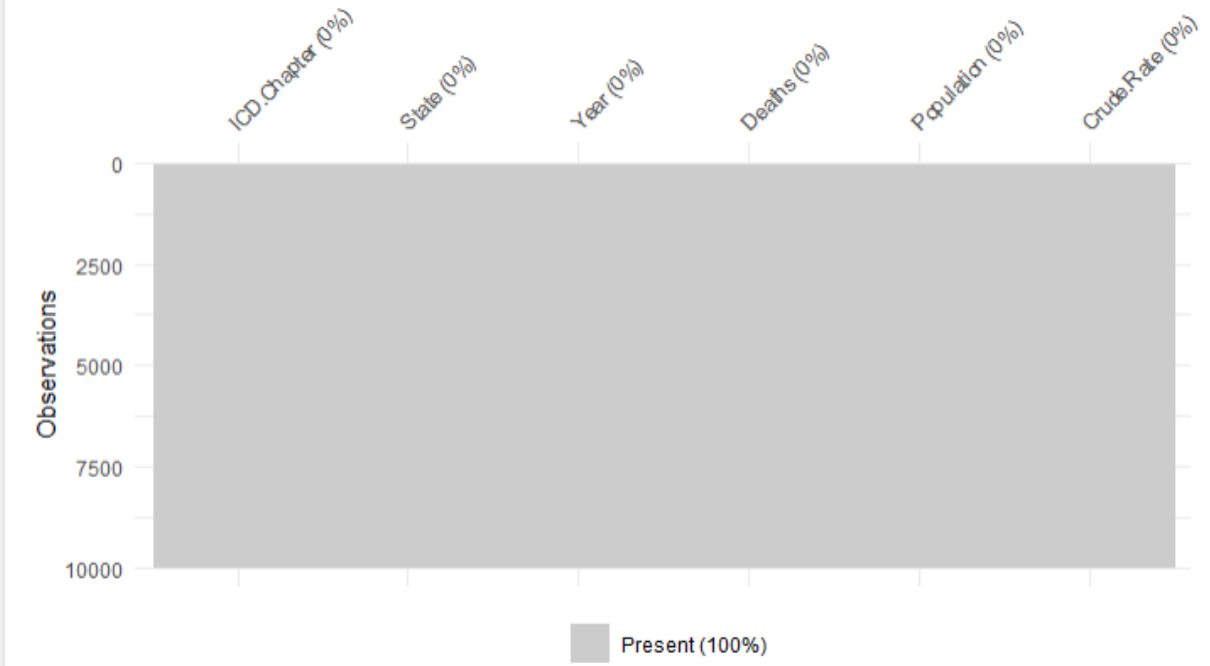
```
Observations: 9,961
Variables: 6
$ ICD.Chapter <fct> Certain infectious and parasitic diseases, Certain infectious and parasitic diseases, Certain infe...
$ State       <fct> AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AL, AK, AK, AK, AK, AK, AK, AK, AK, AK, AK, AK, AZ...
$ Year        <int> 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 1999, 2000, 2001, 2002, 20...
$ Deaths      <int> 1092, 1188, 1211, 1215, 1350, 1251, 1303, 1312, 1241, 1385, 1381, 1358, 61, 51, 58, 67, 61, 53, 68...
$ Population  <int> 4430141, 4447100, 4467634, 4480089, 4503491, 4530729, 4569805, 4628981, 4672840, 4718206, 4757938,...
$ Crude.Rate  <dbl> 24.6, 26.7, 27.1, 27.1, 30.0, 27.6, 28.5, 28.3, 26.6, 29.4, 29.0, 28.4, 9.8, 8.1, 9.2, 10.4, 9.4, ...
```

```{r}
summary(cdc)
```

```
                                                          ICD.Chapter       State           Year          Deaths      
 Certain conditions originating in the perinatal period                 : 612   CA     : 209   Min.   :1999   Min.   :   10  
 Certain infectious and parasitic diseases                              : 612   TX     : 207   1st Qu.:2002   1st Qu.:  177  
 Congenital malformations, deformations and chromosomal abnormalities   : 612   FL     : 205   Median :2005   Median :  667  
 Diseases of the circulatory system                                     : 612   IL     : 205   Mean   :2005   Mean   : 2929  
 Diseases of the digestive system                                       : 612   NY     : 205   3rd Qu.:2008   3rd Qu.: 2474  
 Diseases of the genitourinary system                                   : 612   GA     : 204   Max.   :2010   Max.   :96511  
 (Other)                                                                :6289   (Other):8726                                 
   Population         Crude.Rate    
 Min.   :  491780   Min.   :  0.00  
 1st Qu.: 1728292   1st Qu.:  4.60  
 Median : 4219239   Median : 24.00  
 Mean   : 5937896   Mean   : 52.15  
 3rd Qu.: 6562231   3rd Qu.: 50.50  
 Max.   :37253956   Max.   :478.40  
```

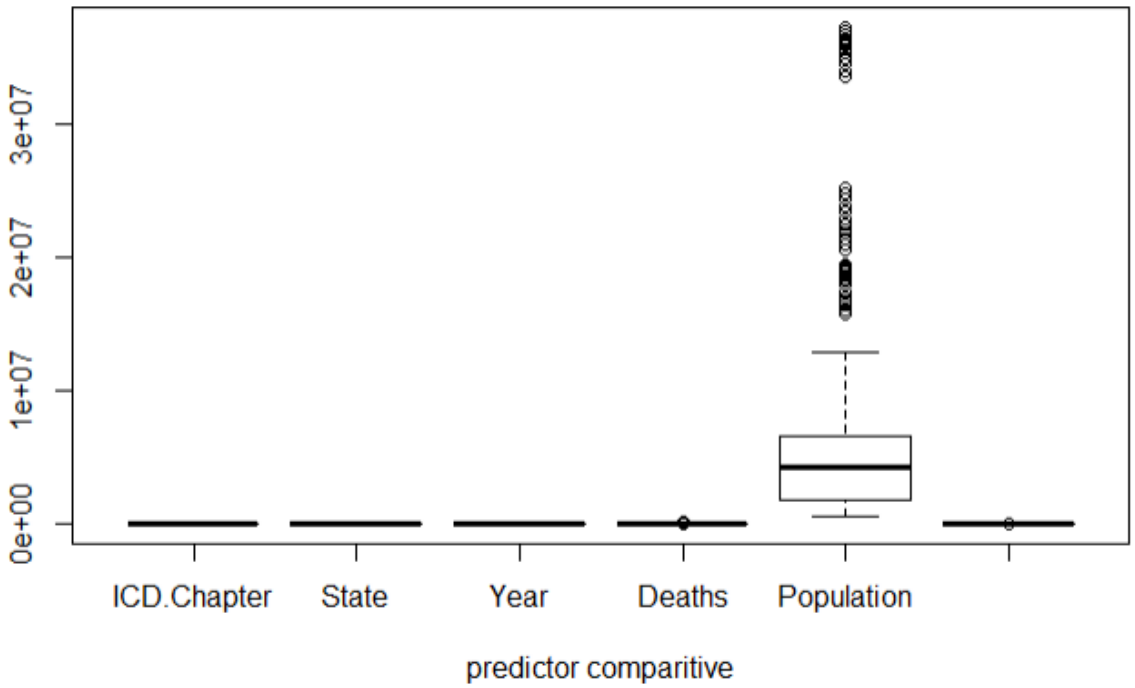
No missing data

```r
vis_miss(cdc)
```



The above plot shows there is no missing data in this database.

```r
boxplot(cdc,xlab="predictor comparitive")
```



```{r}
head(cdc, 100)
```

	ICD.Chapter <fctr>	State <fctr>	Year <int>	Deaths <int>	Population <int>	Crude.Rate <dbl>
1	Certain infectious and parasitic diseases	AL	1999	1092	4430141	24.6
2	Certain infectious and parasitic diseases	AL	2000	1188	4447100	26.7
3	Certain infectious and parasitic diseases	AL	2001	1211	4467634	27.1
4	Certain infectious and parasitic diseases	AL	2002	1215	4480089	27.1
5	Certain infectious and parasitic diseases	AL	2003	1350	4503491	30.0
6	Certain infectious and parasitic diseases	AL	2004	1251	4530729	27.6
7	Certain infectious and parasitic diseases	AL	2005	1303	4569805	28.5
8	Certain infectious and parasitic diseases	AL	2006	1312	4628981	28.3
9	Certain infectious and parasitic diseases	AL	2007	1241	4672840	26.6
10	Certain infectious and parasitic diseases	AL	2008	1385	4718206	29.4

1-10 of 100 rows

Previous 1 2 3 4 5 6 ... 10 Next

After briefly review the data, I feel that the data is clean and ready to be used for further analysis.

Question 1:

- As a researcher, you frequently compare mortality rates from particular causes across different States. You need a visualization that will let you see (for 2010 only) the crude mortality rate, across all States, from one cause (for example, Neoplasms, which are effectively cancers). Create a visualization that allows you to rank States by crude mortality for each cause of death.

```r
library(shiny)
library(forcats)
library(ggplot2)
library(dplyr)

cdc <- read.csv("https://raw.githubusercontent.com/johnpannyc/Jun-Pan-DATA-608-Project-3/master/cleaned-cdc-mortality-1999-2010-2.

#Define UI for application


ui <- fluidPage(

  # Create a title for the panel
  titlePanel("Mortality Rates Across State in USA"),

  # Sidebar layout with input and output definitions
  sidebarLayout(

    # Define input
    sidebarPanel(
      selectInput("cause", "Select a Cause:",
                  sort(unique(cdc$ICD.Chapter))),

      helpText("year 2010")
    ),

    # Create Output
    mainPanel(
      plotOutput("Plot")
    )

  )
)
```

```r
#Define server function required to create the barplot.
server <- function(input, output){



  output$Plot <- renderPlot({

    cdc_data <- cdc %>%
      filter(Year == 2010) %>%
      filter(ICD.Chapter == input$cause) %>%
      select(c(ICD.Chapter,State,Crude.Rate))



    cdc_data %>%
      mutate(State = fct_reorder(State, Crude.Rate)) %>%
      ggplot( aes(x=State, y=Crude.Rate,width=.5)) +
      geom_bar(stat="identity",position="identity") +
      geom_text(size = 5, aes(label = Crude.Rate), position = position_dodge(width = 1),
                inherit.aes = TRUE,
                hjust = -0.5) +
      coord_flip()}, height = 800, width = 750)
}




#Create the shiny app object
shinyApp(ui = ui, server = server)
```

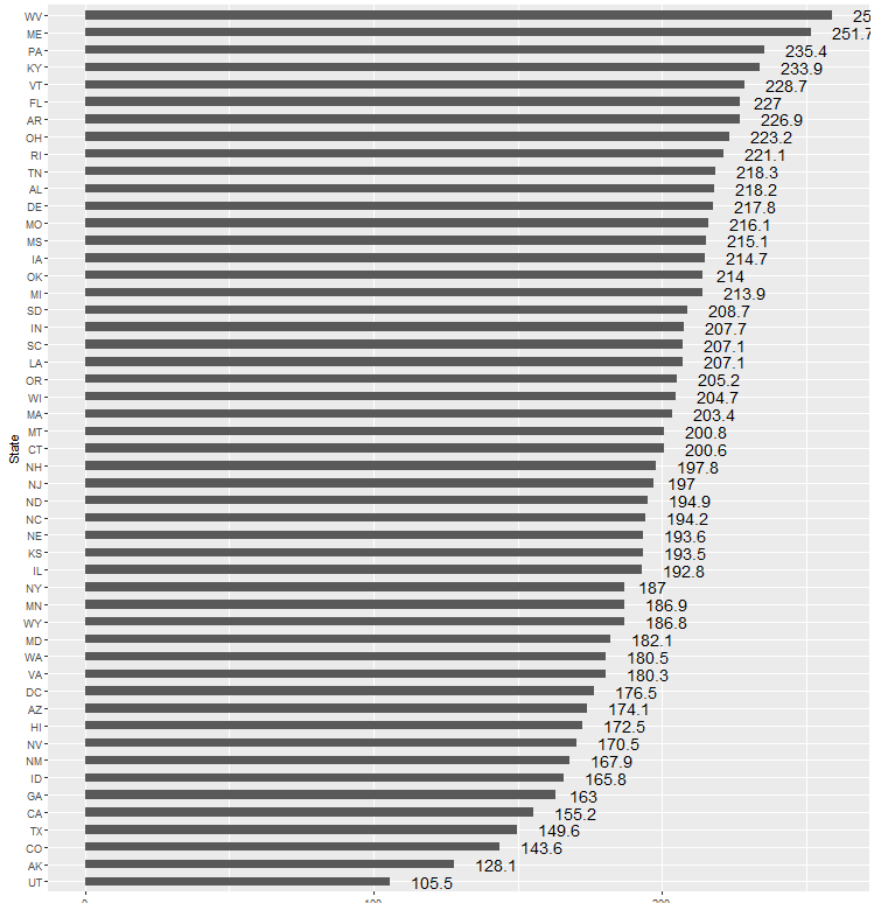
Mortality Rates of Neoplasm in 2010

Mortality Rates Across State in USA

Select a Cause:

Neoplasms

year 2010



Question 2:

- Often you are asked whether particular States are improving their mortality rates (per cause) faster than, or slower than, the national average. Create a visualization that lets your clients see this for themselves for one cause of death at the time. Keep in mind that the national average should be weighted by the national population.

```r
library(shiny)
library(ggplot2)
library(dplyr)
library(forcats)

cdc <-read.csv("https://raw.githubusercontent.com/johnpannyc/Jun-Pan-DATA-608-Project-3/master/cleaned-cdc-mortality-1999-2010-2.

#Define UI
ui <- fluidPage(

  # Title of Panel
  titlePanel("Mortality Rate by Year: black line = national average"),

  # Sidebar Layout
  sidebarLayout(

    # Define input
    sidebarPanel(
      selectInput("cause", "Select a Cause:",
                  choices=as.character(unique(cdc$ICD.Chapter))),
      selectInput("state", "Select a State:",
                  choices=as.character(unique(cdc$State)))

    ),

    # Define output
    mainPanel(
      plotOutput("Plot")
    )

  )
)
```

```r
#Define Server function to create bar plot
server <- function(input, output){


  output$Plot <- renderPlot({

    clean_data <- cdc %>%
      filter(ICD.Chapter == input$cause) %>%
      group_by(Year) %>%
      mutate(weight=(Population/sum(Population))*Crude.Rate) %>%
      mutate(avg=sum(weight)) %>%
      filter(State == input$state)



    clean_data %>%
      ggplot( aes(x=Year, y=Crude.Rate,width=.5)) +
      geom_bar(stat="identity",position="identity",fill="grey", alpha = 0.7) +
      geom_text(size = 5, aes(label = Crude.Rate), position = position_dodge(width = 1),
                inherit.aes = TRUE,
                hjust = 0.5) +
      geom_line(aes(x=Year, y=avg)) +
      theme(axis.text=element_text(size=10),
            axis.title=element_text(size=12))}, height = 800, width = 750)

}


#Create a shiny app object
shinyApp(ui = ui, server = server)
```

Mortality Rate of Neoplasm is higher in Alabama than National Average

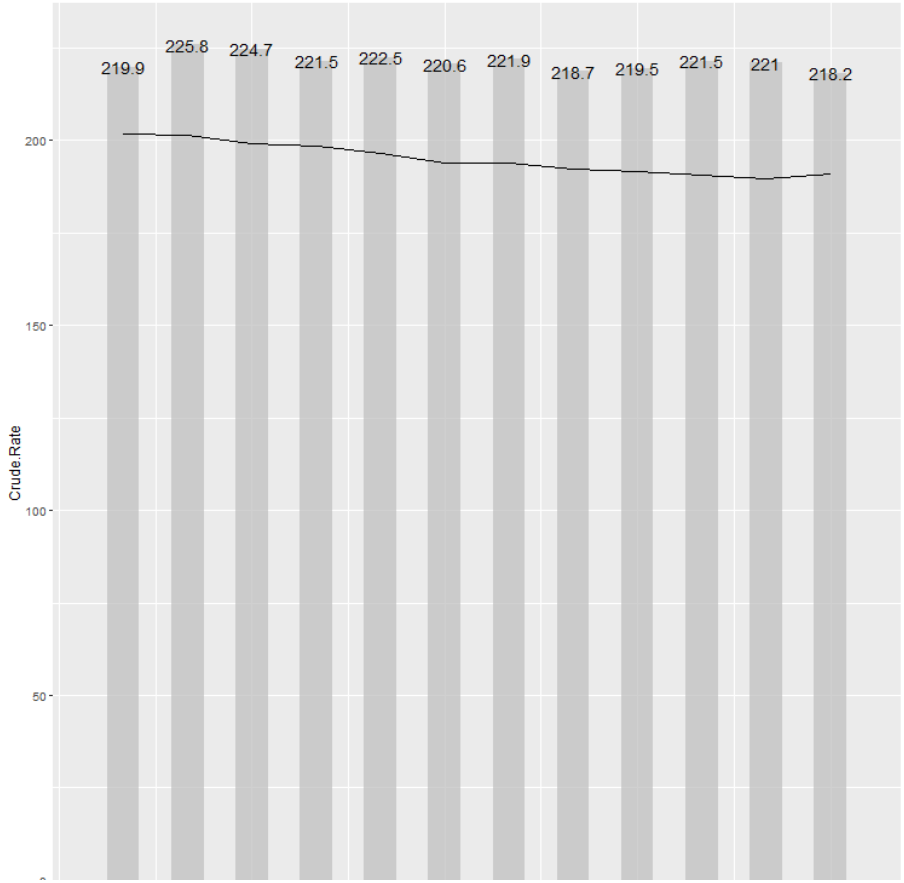
Mortality Rate by Year: black line = national average

Select a Cause:

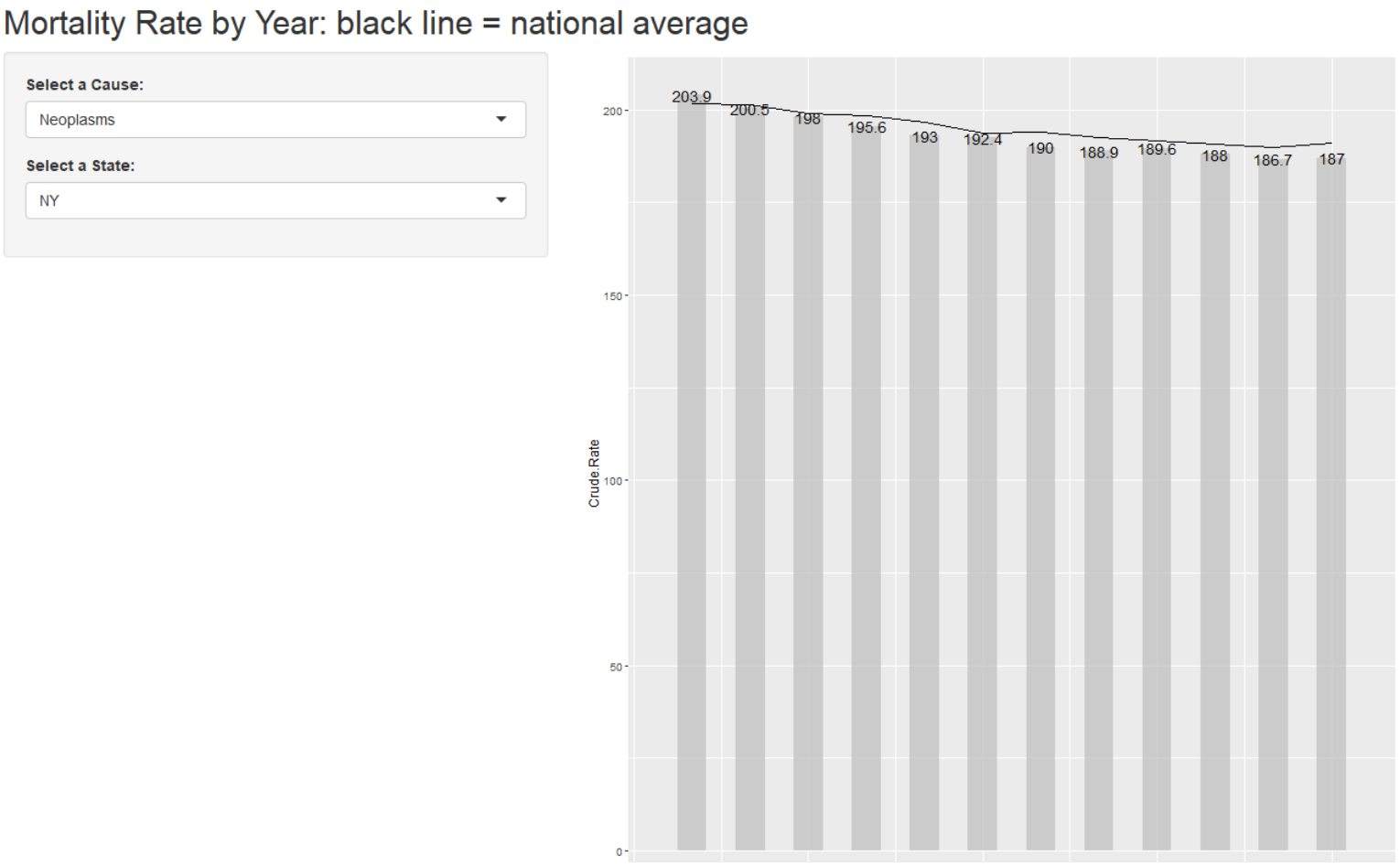
Neoplasms

Select a State:

AL



Mortality Rate of Neoplasm is Close to National Average in New York



Mortality Rate of Neoplasm is lower in Utah Compared to National Average