STA445 - Assignment 8

Richard McCormick

2023-11-14

1. (6 points) At the Insurance Institute for Highway Safety, they have data about human fatalities in vehicle crashes. From this web page, import the data from the Fatal Crash Totals data table and produce a bar graph gives the number of deaths per 100,000 individuals. Be sure to sort the states by highest to lowest mortality. Hint: If you have a problem with the graph being too squished vertically, you can set the chunk options fig.height or fig.width to make the graph larger, but keeping the font sizes the same. The result is that the text is more spread apart. The chunk options out.height and out.width shrink or expand everything in the plot. By making the fix.XXX options large and out.XXX options small, you are effectively decreasing the font size of all the elements in the graph. The other trick is to reset the font size using a theme element_text option: theme(text = element_text(size = 9)).

```
url = 'https://www.iihs.org/topics/fatality-statistics/detail/state-by-state'
page <- read_html(url)
data <- page %>%
  html_nodes('table') %>%
  .[[1]] %>%
  html_table(header=TRUE, fill=FALSE)
```

```
## Warning: The ‘fill‘ argument of ‘html_table()‘ is deprecated as of rvest 1.0.0.
## i An improved algorithm fills by default so it is no longer needed.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

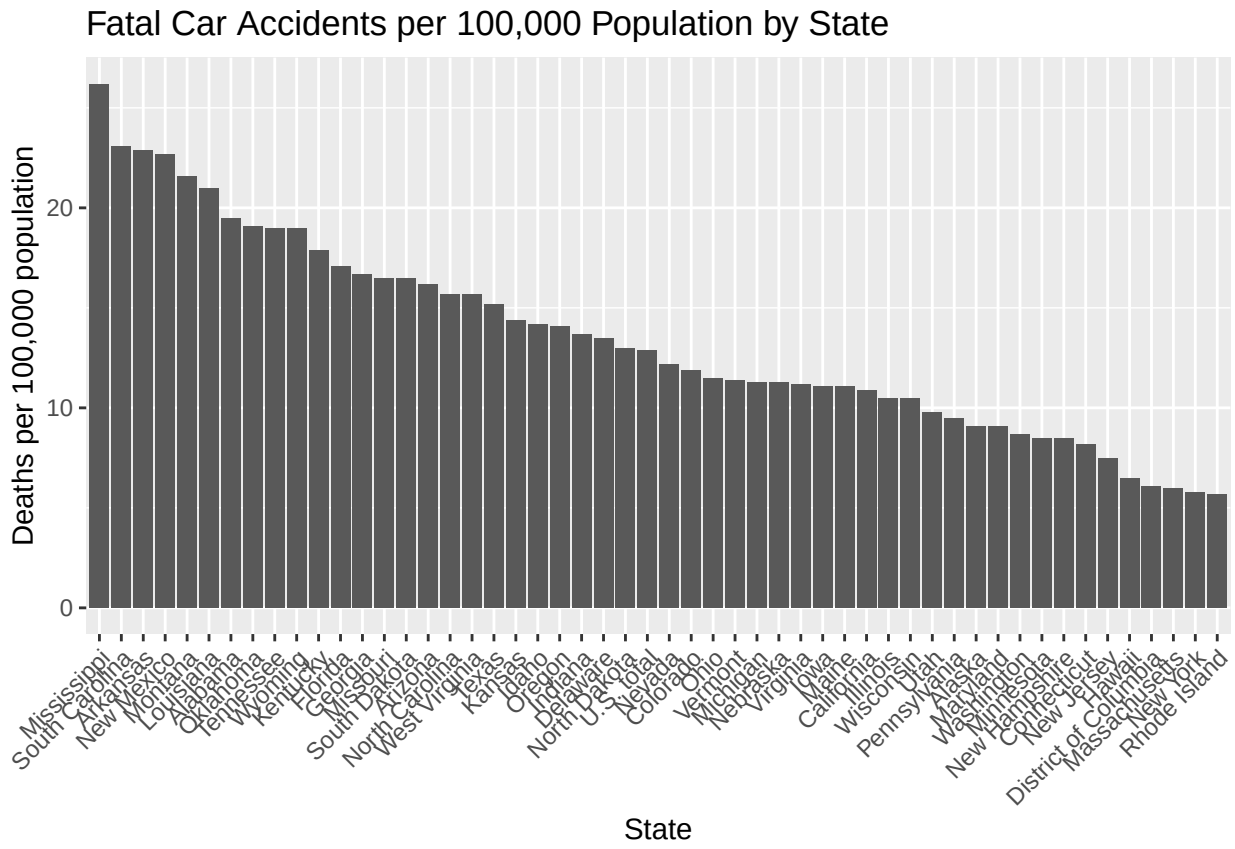
```
names(data) <- data[1,]
```

```
## Warning: The ‘value‘ argument of ‘names<-‘ must be a character vector as of tibble
## 3.0.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
data <- data[-1,]

data$`Deaths per 100,000 population` <- as.numeric( gsub( ",", "", data$`Deaths per 100,000 population`
```

```
ggplot( data=data, aes( x=reorder( State, -`Deaths per 100,000 population` ),
                        y=`Deaths per 100,000 population` ) ) +
  geom_bar( stat='identity' ) +
  theme(axis.text.x = element_text(angle = 45, hjust=1) ) +
  labs( title="Fatal Car Accidents per 100,000 Population by State",
        x="State" )
```



2. (7 points) From the same IIHS website, import the data about seat belt use. Join the Fatality data with the seat belt use and make a scatter plot of percent seat belt use vs number of fatalities per 100,000 people.

```
belt.data <- page %>%
  html_nodes('table') %>%
  .[[5]] %>%
  html_table(header=TRUE, fill=FALSE)

names(belt.data) <- belt.data[1,]
belt.data <- belt.data[-1,]
belt.data = belt.data[-1,]

names(belt.data) <- c("State",
                      "Percent of Observed Seat Belt Use",
```

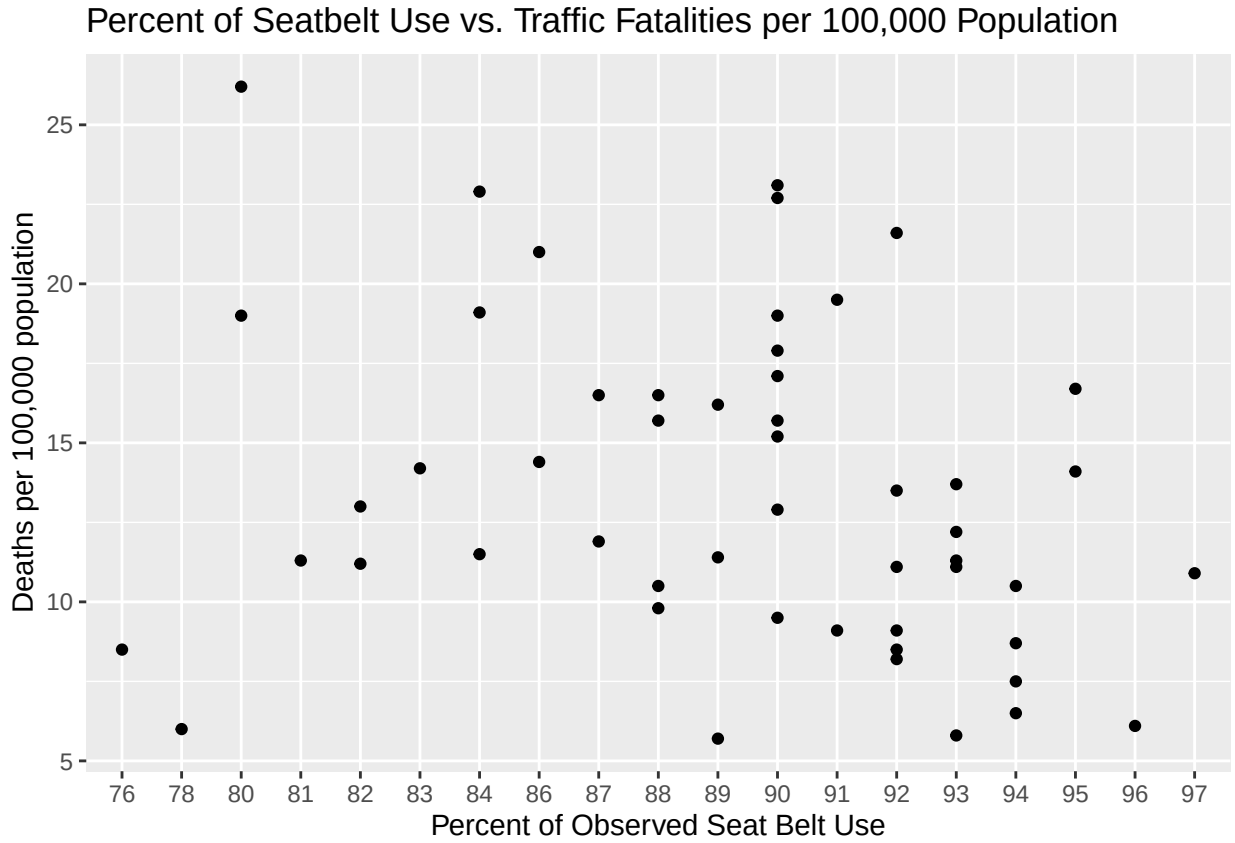
```
                      "Restrined Fatally Injured Occupants",
                      "Percent of Restrained Fatally Injured Occupants",
                      "")
```

```
## Warning: The ‘value‘ argument of ‘names<-‘ must have the same length as ‘x‘ as of tibble
## 3.0.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## Warning: The ‘value‘ argument of ‘names<-‘ can’t be empty as of tibble 3.0.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
data <- merge( data, belt.data )

ggplot( data=data,
        aes( x=`Percent of Observed Seat Belt Use`,
             y=`Deaths per 100,000 population` ) ) +
  geom_point() +
  labs( title="Percent of Seatbelt Use vs. Traffic Fatalities per 100,000 Population" )
```



3. (Skip: Come back to it later if you are interested.) From the NAU sub-reddit, extract the most recent threads.

```
url <- 'https://www.reddit.com/r/NAU/'
page <- read_html(url)

HeadLines <- page %>%
  html_nodes('shreddit-post') %>%    # Grab just the headlines
  html_text()  # Convert the <a>Text</a> to just Text

reddit.data <- data.frame( HeadLines) %>%
  mutate( User = str_extract( HeadLines, "u/\\w*") ) %>%
  mutate( strings = strsplit(as.character(HeadLines), "\\n") )

i <- 1

for (string in reddit.data$strings)
{
  new_str <- str_trim( string )
  new_str <- new_str[new_str != ""]

  reddit.data[i, "Title"] <- new_str[3]
  reddit.data[i, "Post"] <- new_str[4]

  i <- i + 1
}

rownames( reddit.data ) <- NULL
reddit.data$HeadLines <- NULL
reddit.data$strings <- NULL


print( reddit.data )
```

```
##              User
## 1  u/GennaroIsGod
## 2       u/Revomby
## 3 u/ThomasTheToad
##                                                                  Title
## 1                                                  NAU Discord Server
## 2 Looking for Someone to Finish the Remainder of my Lease (Urgent)
## 3          Has anyone else been having issues with on-campus wifi?
##
## 1
## 2 I am giving up my lease at Yugo Flagstaff Central and need someone to take over starting January 2
## 3
```

```
print( reddit.data$Post )
```

```
## [1] "https://www.reddit.com/r/NAU/comments/pf9vb1/nau_discord_server/"
## [2] "I am giving up my lease at Yugo Flagstaff Central and need someone to take over starting January
## [3] "It's been cutting out randomly throughout the day for me no matter where I am on campus. I'm wo
```