Module16

Joe Vargovich

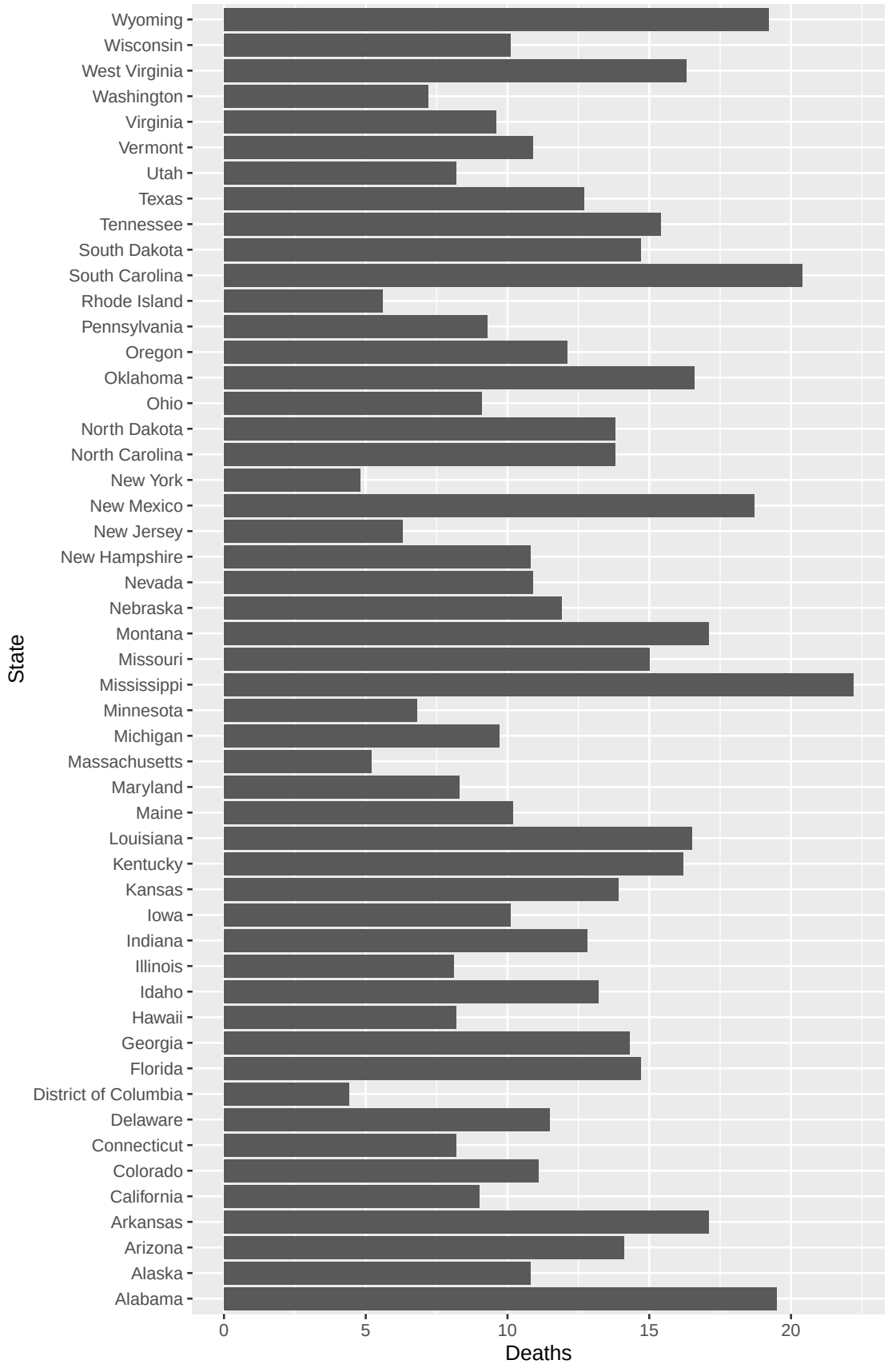
11/5/2020

#Exercise 1 - Crash Data Scrape

```
url = "https://www.iihs.org/topics/fatality-statistics/detail/state-by-state"
page = read_html(url)

CrashData = page %>%
  html_nodes('table') %>%
  .[[1]] %>%
  html_table(header=FALSE, fill=TRUE) %>%
  select(X1,X2, X6) %>%
  slice(-1:-2) %>%
  magrittr::set_colnames(c('State','Population', 'Deaths')) %>%
  mutate_at( vars(matches('Deaths')), str_remove_all, ',') %>%           # remove all commas
  mutate_at( vars(matches('Deaths')), str_remove, '\\[[0-9]+\\]') %>%    # remove [7] stuff
  mutate_at( vars( matches('Deaths')), as.numeric) %>%
  mutate_at( vars(matches('Population')), str_remove_all, ',') %>%           # remove all commas
  mutate_at( vars(matches('Population')), str_remove, '\\[[0-9]+\\]') %>%    # remove [7] stuff
  mutate_at( vars( matches('Population')), as.numeric) %>%
  filter(State != "U.S. total")
head(CrashData)
```

```
##        State Population Deaths
## 1    Alabama    4887871   19.5
## 2     Alaska     737438   10.8
## 3    Arizona    7171646   14.1
## 4   Arkansas    3013825   17.1
## 5 California   39557045    9.0
## 6   Colorado    5695564   11.1
```

```
#Plot our data
plotOfDeath = ggplot(CrashData, aes(x=State, y=Deaths)) +
  geom_col() +
  coord_flip()
plotOfDeath
```


Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Georgia
Florida
District of Columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alaska
Alabama
State
0
5
10
15
20
Deaths

#Exercise

2 -

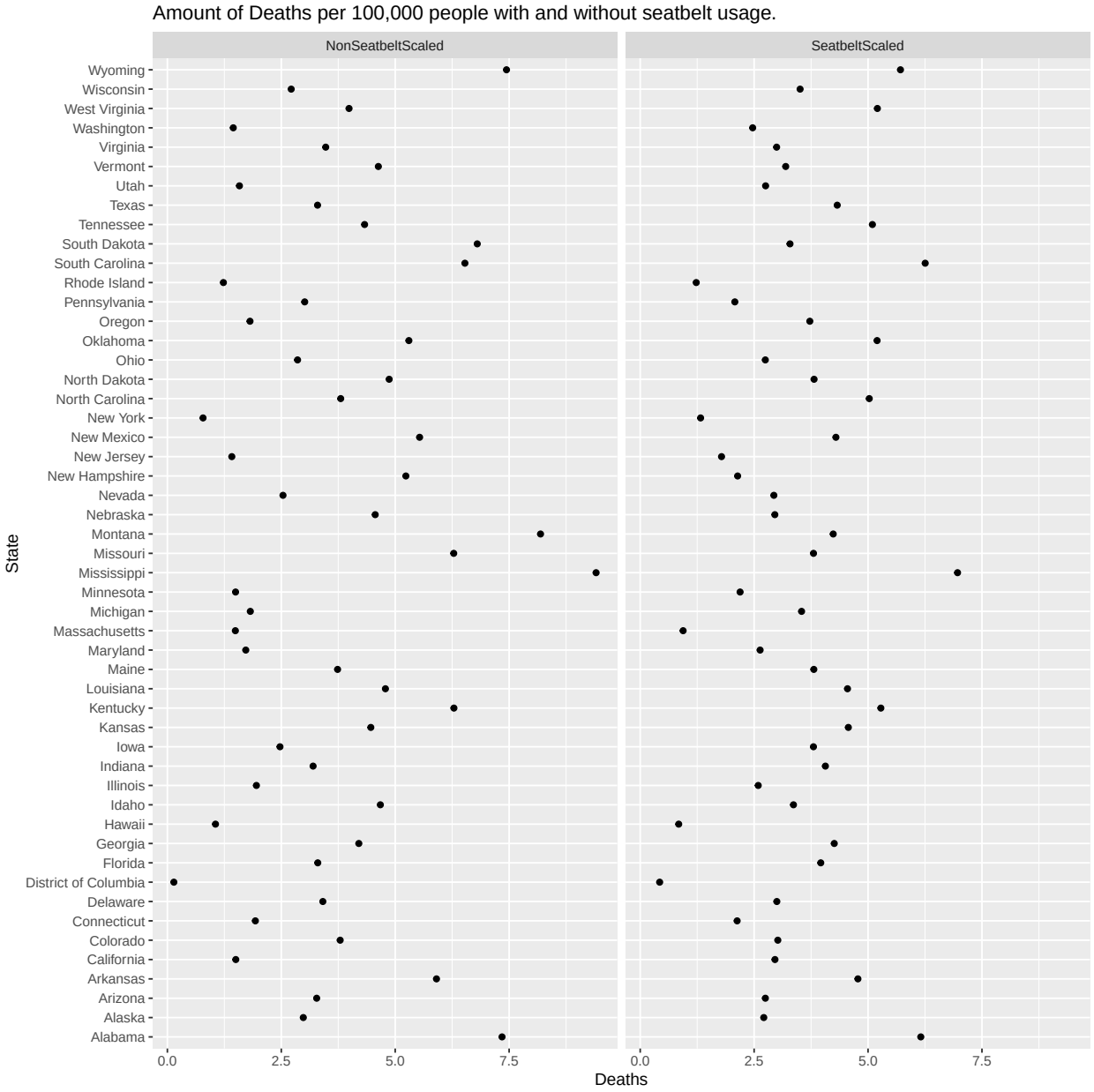
```
url = "https://www.iihs.org/topics/fatality-statistics/detail/state-by-state"
page = read_html(url)

CrashDataRestraint = page %>%
  html_nodes('table') %>%
  .[[5]] %>%
  html_table(header=FALSE, fill=TRUE) %>%
  select(X1, X3, X5) %>%
  slice(-1:-3) %>%
  magrittr::set_colnames(c('State', 'SeatbeltDeaths', 'NonSeatbeltDeaths')) %>%
  mutate_at( vars(matches('SeatbeltDeaths')), str_remove_all, ',') %>%           # remove all commas
  mutate_at( vars(matches('SeatbeltDeaths')), str_remove, '\\[[0-9]+\\]') %>%    # remove [7] stuff
  mutate_at( vars( matches('SeatbeltDeaths')), as.numeric) %>%
  mutate_at( vars(matches('NonSeatbeltDeaths')), str_remove_all, ',') %>%           # remove all commas
  mutate_at( vars(matches('NonSeatbeltDeaths')), str_remove, '\\[[0-9]+\\]') %>%    # remove [7] stuff
  mutate_at( vars( matches('NonSeatbeltDeaths')), as.numeric)
joinedDf = inner_join(CrashData, CrashDataRestraint)
```

```
## Joining, by = "State"
```

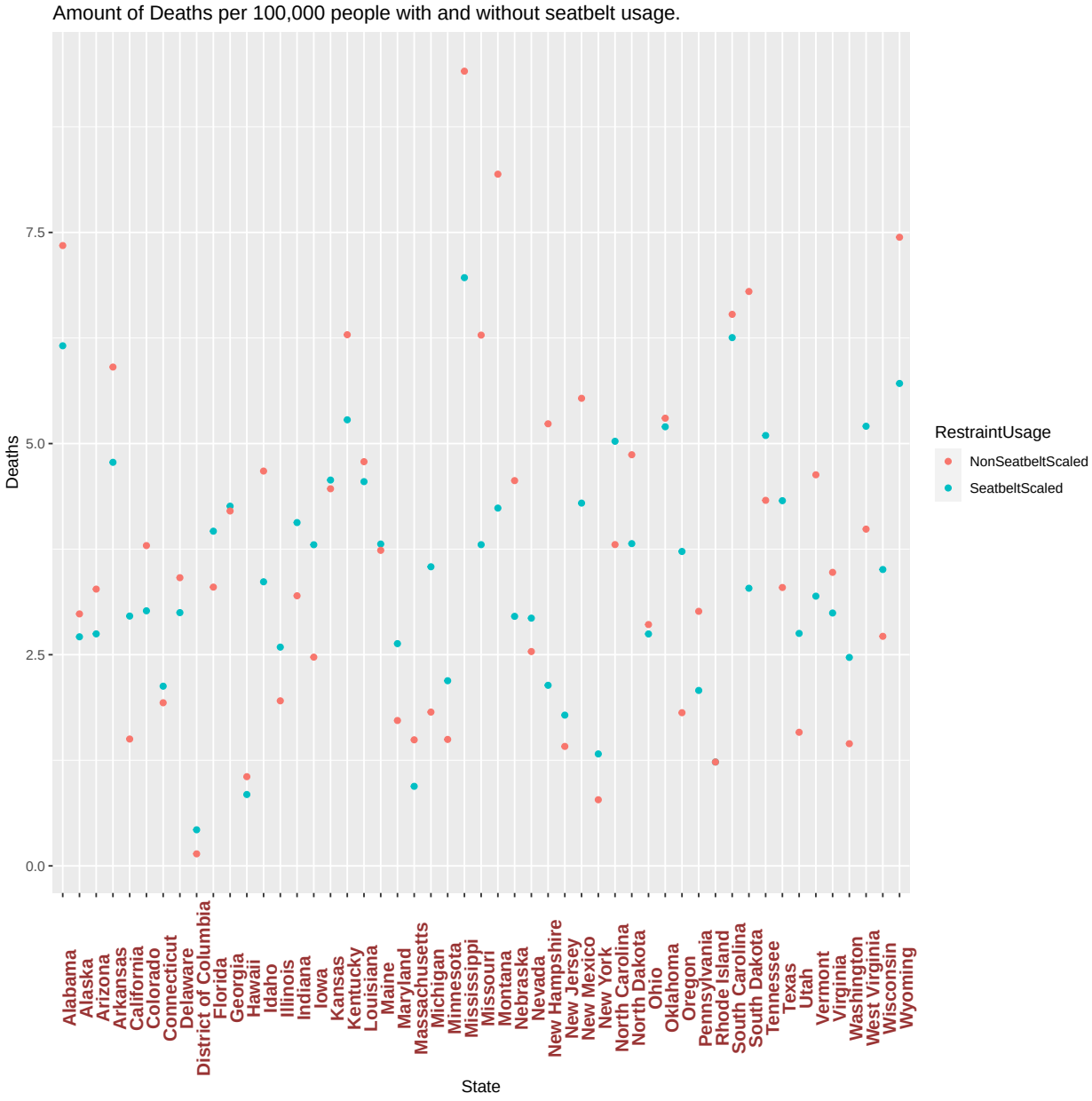
```
joinedDf = joinedDf %>%
  mutate(
    SeatbeltScaled = SeatbeltDeaths/Population * 100000,
    NonSeatbeltScaled = NonSeatbeltDeaths/Population * 100000
  ) %>%
  select(State, SeatbeltScaled, NonSeatbeltScaled) %>%
  pivot_longer(SeatbeltScaled:NonSeatbeltScaled,
               names_to = "RestraintUsage",
               values_to = "Deaths") %>%
  mutate(
    RestraintUsage = factor(RestraintUsage)
    )
head(joinedDf)
```

```
## # A tibble: 6 x 3
##   State   RestraintUsage    Deaths
##   <chr>   <fct>              <dbl>
## 1 Alabama SeatbeltScaled      6.16
## 2 Alabama NonSeatbeltScaled   7.34
## 3 Alaska  SeatbeltScaled      2.71
## 4 Alaska  NonSeatbeltScaled   2.98
## 5 Arizona SeatbeltScaled      2.75
## 6 Arizona NonSeatbeltScaled   3.28
```

```
#I wasn't sure of the best way to show this data in a scatter plot,
# So I made two for good measure.
restraintPlot = ggplot(joinedDf, aes(x=State, y=Deaths)) +
  geom_point() +
  coord_flip() +
  facet_wrap(.~RestraintUsage) +
  labs(
    title = "Amount of Deaths per 100,000 people with and without seatbelt usage."
  )
restraintPlot
```




```
restraintPlot2 = ggplot(joinedDf, aes(x=State, y=Deaths)) +
  geom_point(aes(color=RestraintUsage)) +
  theme(axis.text.x = element_text(face = "bold", color = "#993333",
                           size = 12, angle = 90)) +
  labs(
    title = "Amount of Deaths per 100,000 people with and without seatbelt usage."
  )
restraintPlot2
```



Exercise 3 - Reddit Headlines Data Scraping

```
url = "https://www.reddit.com/r/NAU/new"
page = read_html(url)

firstHeadlines = page %>%
  html_nodes('.SQnoC3ObvgnGjWt90zD9Z ._eYtD2XCVieq6emjKBH3m') %>%
  html_text()

head(firstHeadlines)
```

```
## [1] "where is covid testing taking place?"
## [2] "Winter Term?"
```

```
## [3] "ENG 105-001 critical thinking and writing in the university community. Can I take this as a fres
## [4] "Does the Cline library kick people out after they close?"
## [5] "To those who decided to throw large parties for tequila sunrise"
```