

S01 - US Counties Demonstration

1. Load the packages `tidyverse` and `usdata`.

```
# load packages
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.1 --
## v ggplot2 3.3.5      v purrr  0.3.4
## v tibble  3.1.4      v dplyr  1.0.7
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   2.0.1      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(usdata)
```

2. Open the `county_complete` data set.

```
# apply select() to get three columns
county_sub <- county_complete %>%
  select(pop_2019,unemployment_rate_2019,povertry_2019,state,name)

# glimpse the subsetted data
glimpse(county_sub)

## Rows: 3,142
## Columns: 5
## $ pop_2019          <dbl> 55380, 212830, 25361, 22493, 57681, 10248, 1982~
## $ unemployment_rate_2019 <dbl> 3.5, 4.0, 9.4, 7.0, 3.1, 4.1, 7.0, 7.2, 4.0, 4.~
## $ povertry_2019      <dbl> 15.2, 10.4, 30.7, NA, 13.6, NA, NA, 17.9, 17.3,~
## $ state              <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Al~
## $ name               <chr> "Autauga County", "Baldwin County", "Barbour Co~
```

In our dataset, there are 3,142 rows, which is corresponding to the number of counties in the USA.

3. Answer these following questions

- What are the top 10 US states by 2019 population?
- What is the distribution of unemployment rate in 2019 for counties in California, Oregon, and/or Washington?
- What is the relationship between unemployment rate and poverty rate?

a.

```
county_sub %>%
  # group by state, we know that each state has different number of counties
```

```

group_by(state) %>%
  # take the sum of the population for each state
  summarise(sum = sum(pop_2019)) %>%
  # sort them from highest to lowest
  arrange(desc(sum)) %>%
  # print top 10
  head(10)

```

```

## # A tibble: 10 x 2
##   state      sum
##   <chr>      <dbl>
## 1 California 39283497
## 2 Texas      28260856
## 3 Florida    20901636
## 4 New York   19572319
## 5 Pennsylvania 12791530
## 6 Illinois    12770631
## 7 Ohio        11655397
## 8 Georgia     10403847
## 9 North Carolina 10264876
## 10 Michigan    9965265

```

Therefore, the top 10 most populous state is CA, TX, FL, NY, PA, IL, GA, NC, and MI.

b.

```

county_new <- county_sub %>%
  filter(state == "California" |
         state == "Oregon" |
         state == "Washington")
glimpse(county_new)

```

```

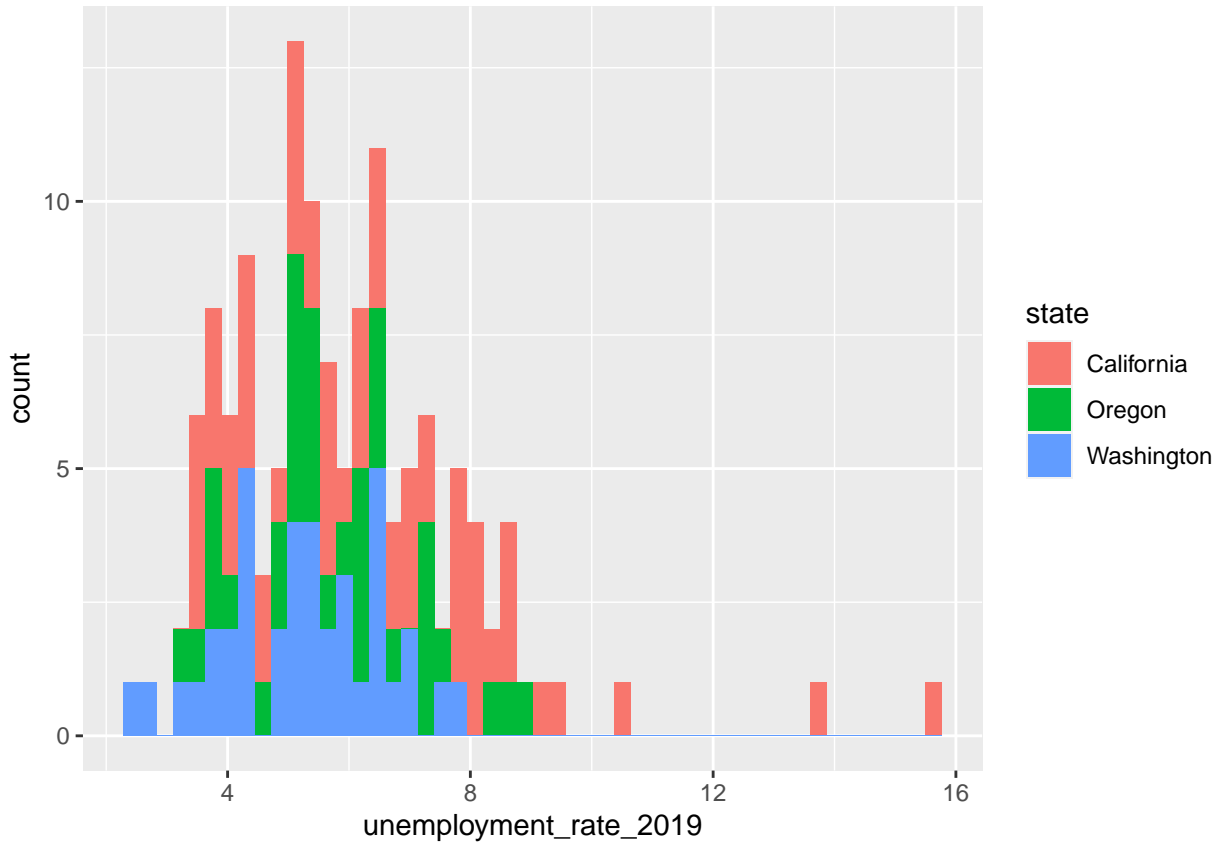
## Rows: 133
## Columns: 5
## $ pop_2019      <dbl> 1656754, 1039, 38429, 225817, 45514, 21454, 114~
## $ unemployment_rate_2019 <dbl> 4.4, 15.6, 6.6, 7.7, 6.6, 4.4, 5.0, 7.8, 5.4, 8~
## $ poverty_2019  <dbl> 9.9, NA, NA, 19.1, NA, NA, 8.7, NA, 8.4, 22.5, ~
## $ state         <chr> "California", "California", "California", "Cali~
## $ name          <chr> "Alameda County", "Alpine County", "Amador Coun~

```

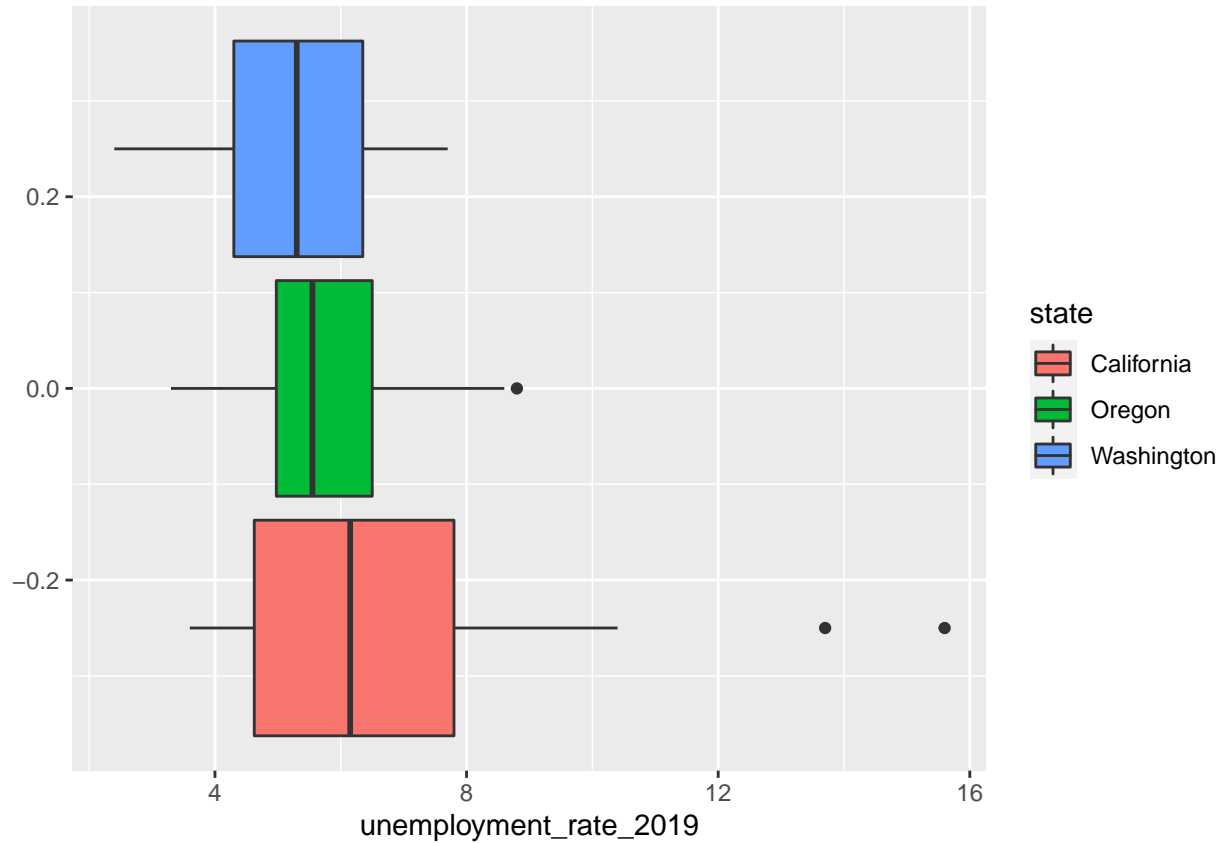
```

ggplot(data = county_new, aes(x = unemployment_rate_2019, fill = state)) + geom_histogram(bins = 50)

```



```
ggplot(data = county_new, aes(x = unemployment_rate_2019, fill = state)) + geom_boxplot()
```

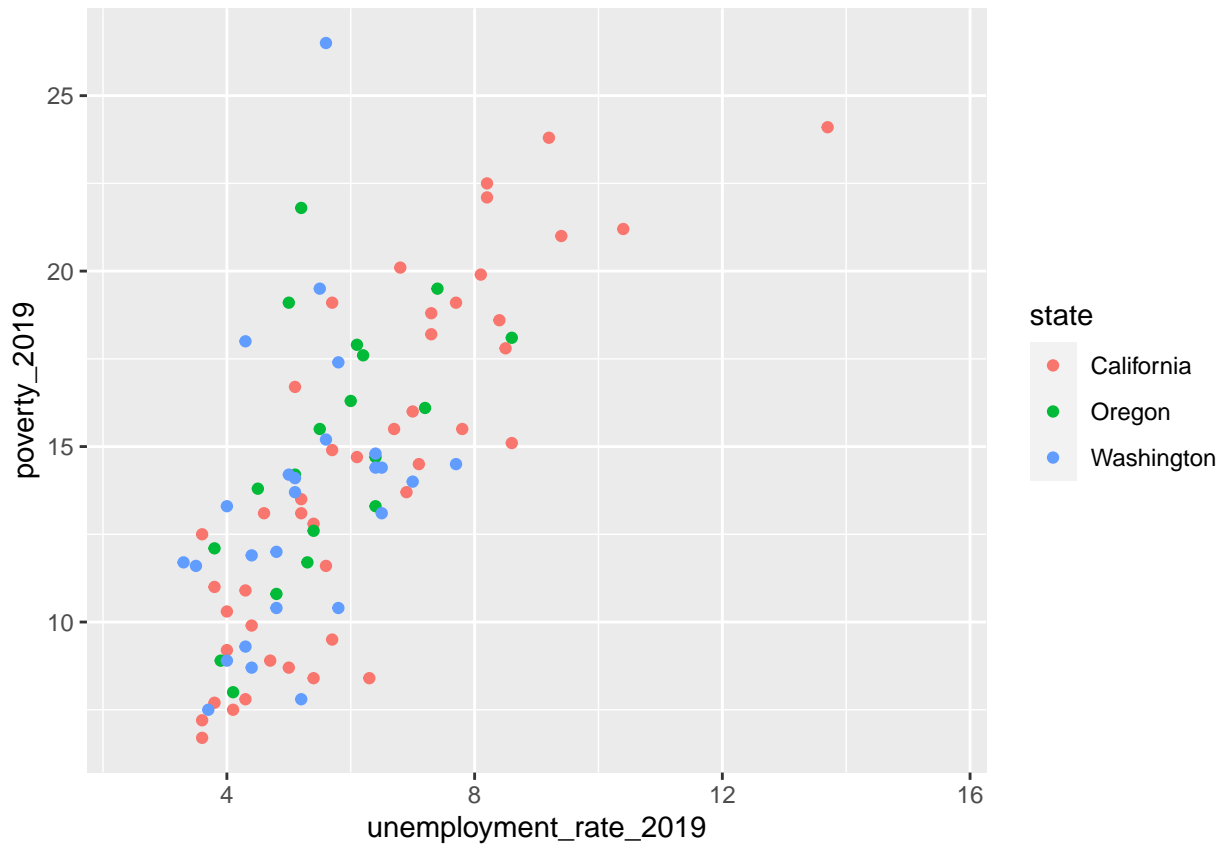


According to the histogram, the distribution for the California counties unemployment rate for 2019 is unimodal and slightly skewed-right. Similarly, the distribution for Oregon and Washington is also unimodal but not as skewed as California. There are some outliers in California and Oregon. Also, California has the highest median unemployment rate in 2019 among the three.

c.

```
ggplot(data = county_new, aes(x = unemployment_rate_2019, y = poverty_2019)) +
  geom_point(aes(color = state))
```

```
## Warning: Removed 44 rows containing missing values (geom_point).
```



All three states have a positive linear association between unemployment and poverty.