

# complete

## Data Cleaning

### Load Assets

```
library(tidyverse)
```

Warning: package 'lubridate' was built under R version 4.4.2

```
— Attaching core tidyverse packages — tidyverse 2.0.0 —
✓ dplyr      1.1.4    ✓ readr      2.1.5
✓ forcats    1.0.0    ✓ stringr    1.5.1
✓ ggplot2    3.5.1    ✓ tibble     3.2.1
✓ lubridate  1.9.4    ✓ tidyr      1.3.1
✓ purrr      1.0.2

— Conflicts — tidyverse_conflicts() —
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readxl)
library(writexl)

full <- read_excel("data/bds_data.xlsx", sheet = "BDS")
patents <- read_excel("data/bds_data.xlsx", sheet = "Patents")
stem_edu <- read_excel("data/S&EDegrees2000-2018.xlsx")
```

New names:

- `2018-All S&E degrees/all higher education degrees (Percent)` -> `2018-All S&E degrees/all higher education degrees (Percent)...58`
- `2018-All S&E degrees/all higher education degrees (Percent)` -> `2018-All S&E degrees/all higher education degrees (Percent)...59`

```
employment_growth <- read_excel("data/bds_data.xlsx", sheet = "Copy of Total Employment Growth")
sector_growth <- read_excel("data/bds_data.xlsx", sheet = "Info Employment Growth")
vc <- read_excel("data/Wide Data Set.xlsx", sheet = "VC #", skip = 1)
```

### Make datasets long

```
patents <- patents |>
  pivot_longer(cols = 2:27, values_to = "Number of Utility Patents", names_to = "Year") |>
```

```
mutate(Year = as.numeric(Year))

stem_edu <- stem_edu |>
  head(51) |>
  pivot_longer(cols = 2:59, values_to = "Number", names_to = "Year") |>
  mutate(Number = as.numeric(Number))

vc <- vc |>
  pivot_longer(cols = 2:27, values_to = "Number of Firms Receiving VC", names_to = "Year") |>
  mutate(Year = as.numeric(Year))
```

## Clean S&E Dataset and Fix Formatting

```
se_degrees <- stem_edu |>
  filter(str_detect(Year, "conferred")) |>
  mutate(Year = str_sub(Year, 1, 4), Year = as.numeric(Year)) |>
  rename("S&E Degrees" = Number)

total_degrees <- stem_edu |>
  filter(str_detect(Year, "All higher")) |>
  mutate(Year = str_sub(Year, 1, 4), Year = as.numeric(Year)) |>
  rename("Total Degrees" = Number)

se_degree_proportion <- stem_edu |>
  filter(str_detect(Year, "All S&E")) |>
  mutate(Year = str_sub(Year, 1, 4), Year = as.numeric(Year)) |>
  distinct() |>
  rename("Proportion of S&E" = Number)
```

## Full, long dataset of all the variables of every state in every year

```
full <- full |>
  left_join(patents, join_by(State, Year)) |>
  left_join(se_degrees, join_by(State, Year)) |>
  left_join(total_degrees, join_by(State, Year)) |>
  left_join(se_degree_proportion, join_by(State, Year)) |>
  left_join(vc, join_by(State, Year))

write_xlsx(full, "data/full_dataset.xlsx")

full <- full |>
  mutate(State = factor(State))
```

## Plots

## Separate plots for every year of states vs selected variable

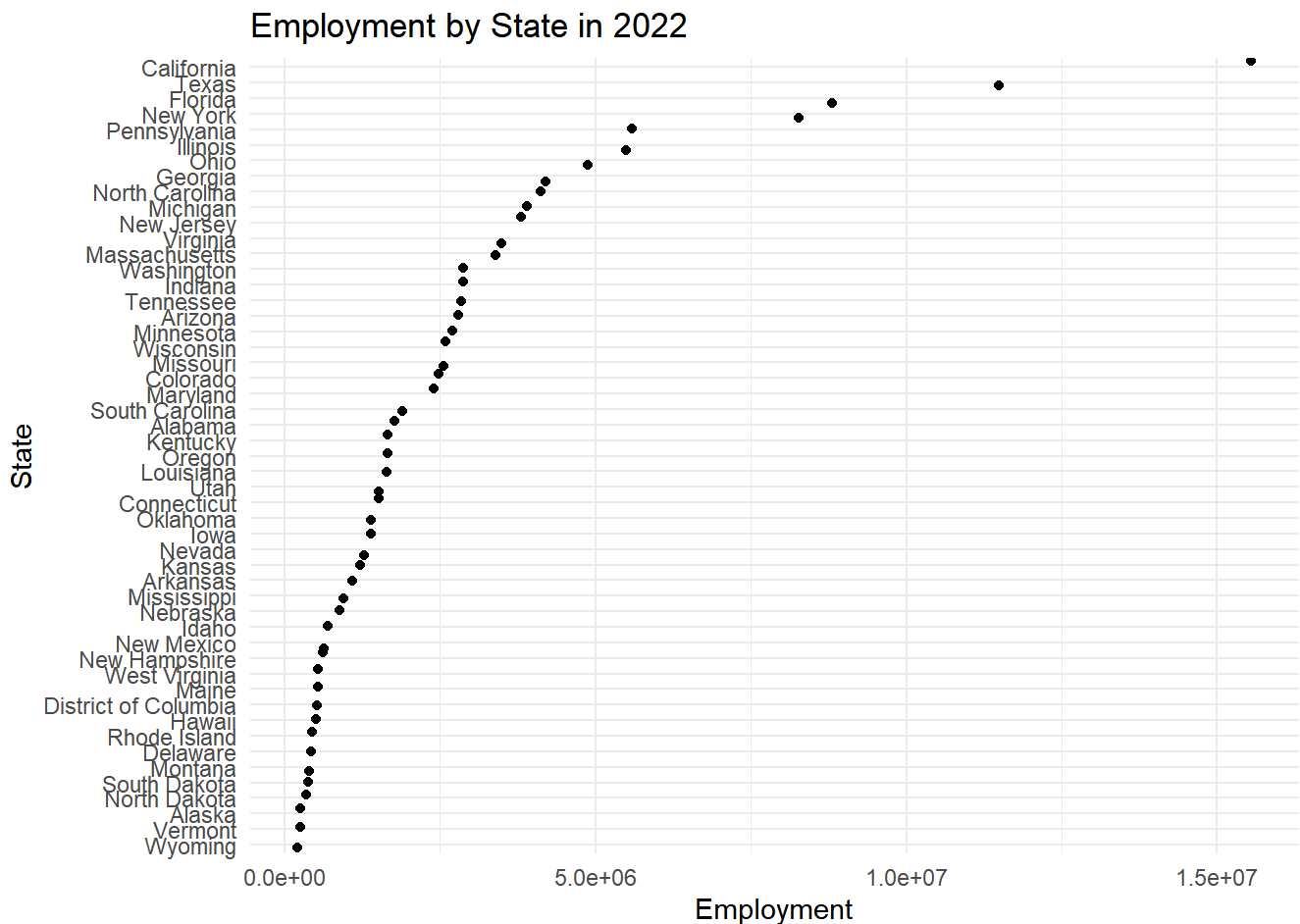
```
years <- unique(full$Year)

for (y in years) {
  full_filtered <- full |>
    filter(Year == y)

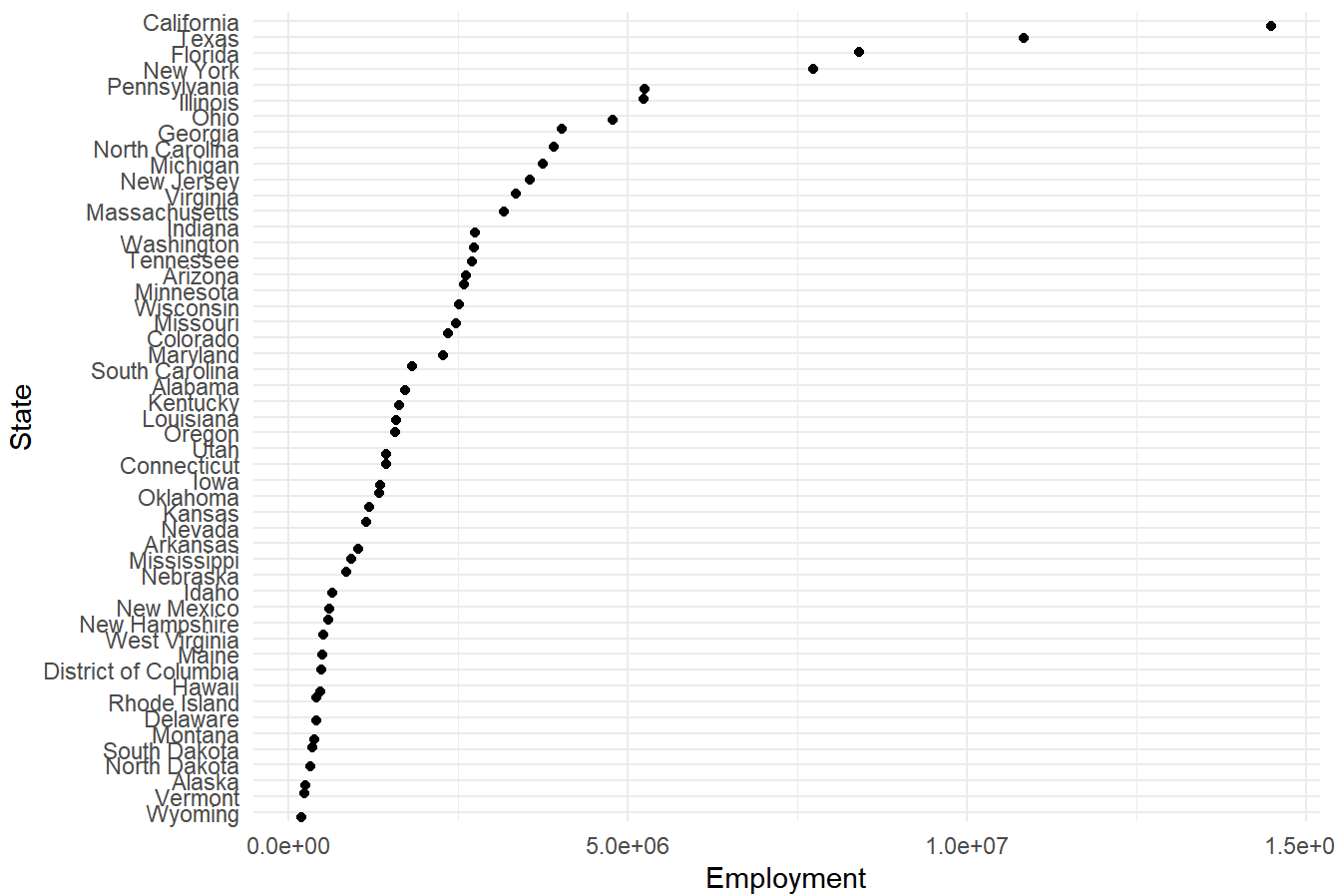
  full_filtered$State <- fct_reorder(full_filtered$State, full_filtered$`Total Employment`)

  plot <- ggplot(full_filtered, aes(x = State, y = `Total Employment`)) +
    geom_jitter() +
    labs(title = paste("Employment by State in", y),
         x = "State",
         y = "Employment") +
    theme_minimal() +
    coord_flip() # Optional for readability

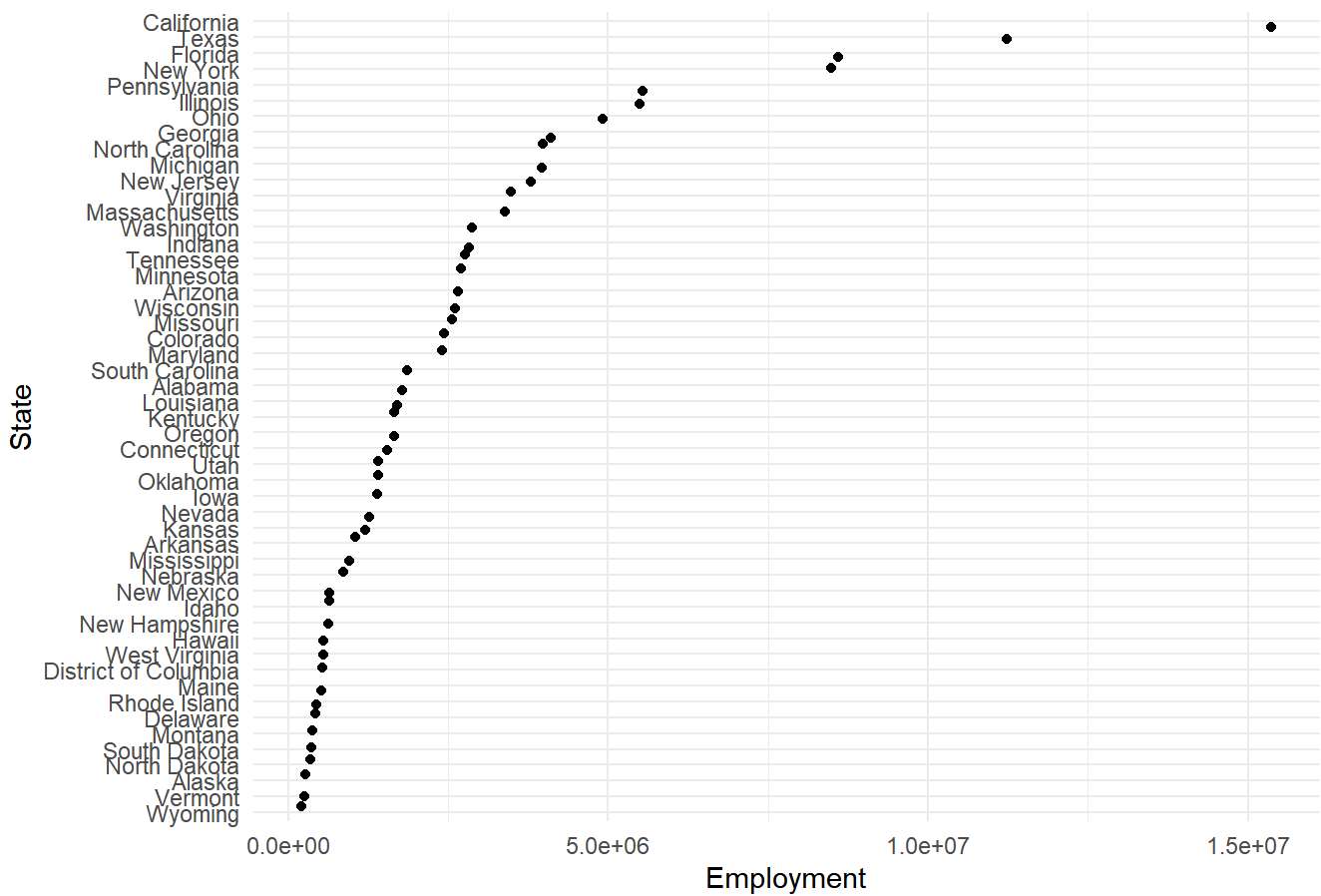
  print(plot)
}
```



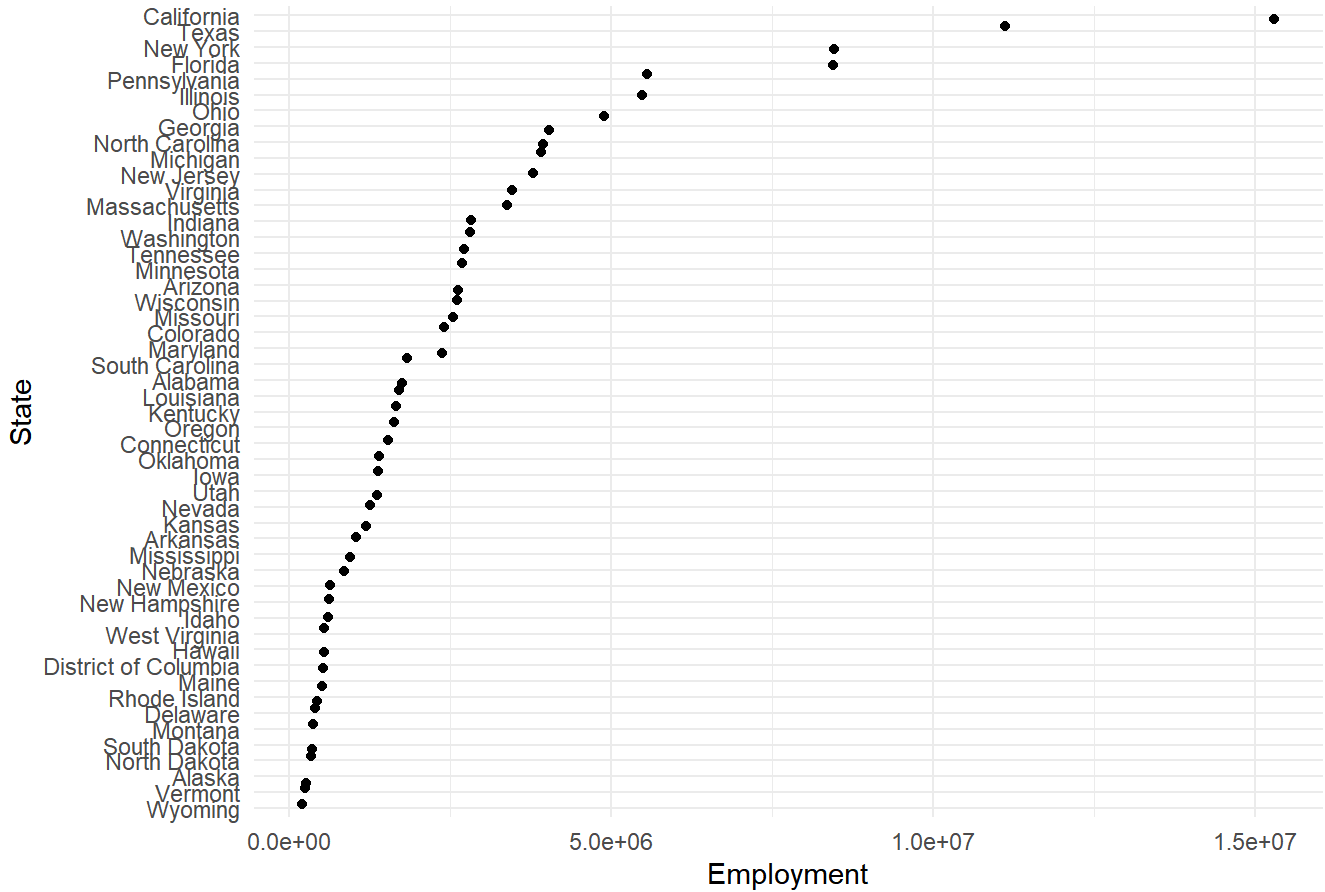
Employment by State in 2021



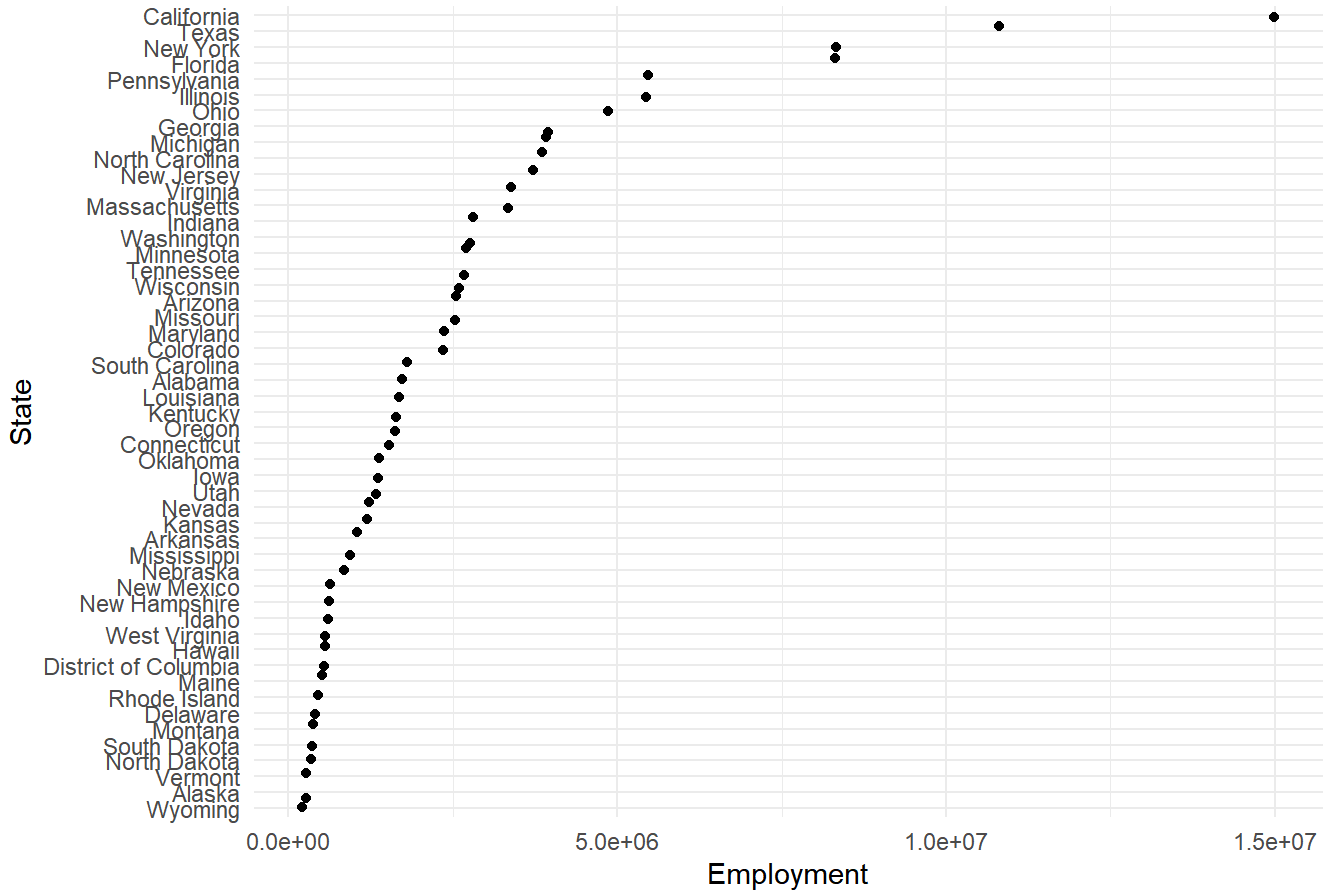
Employment by State in 2020



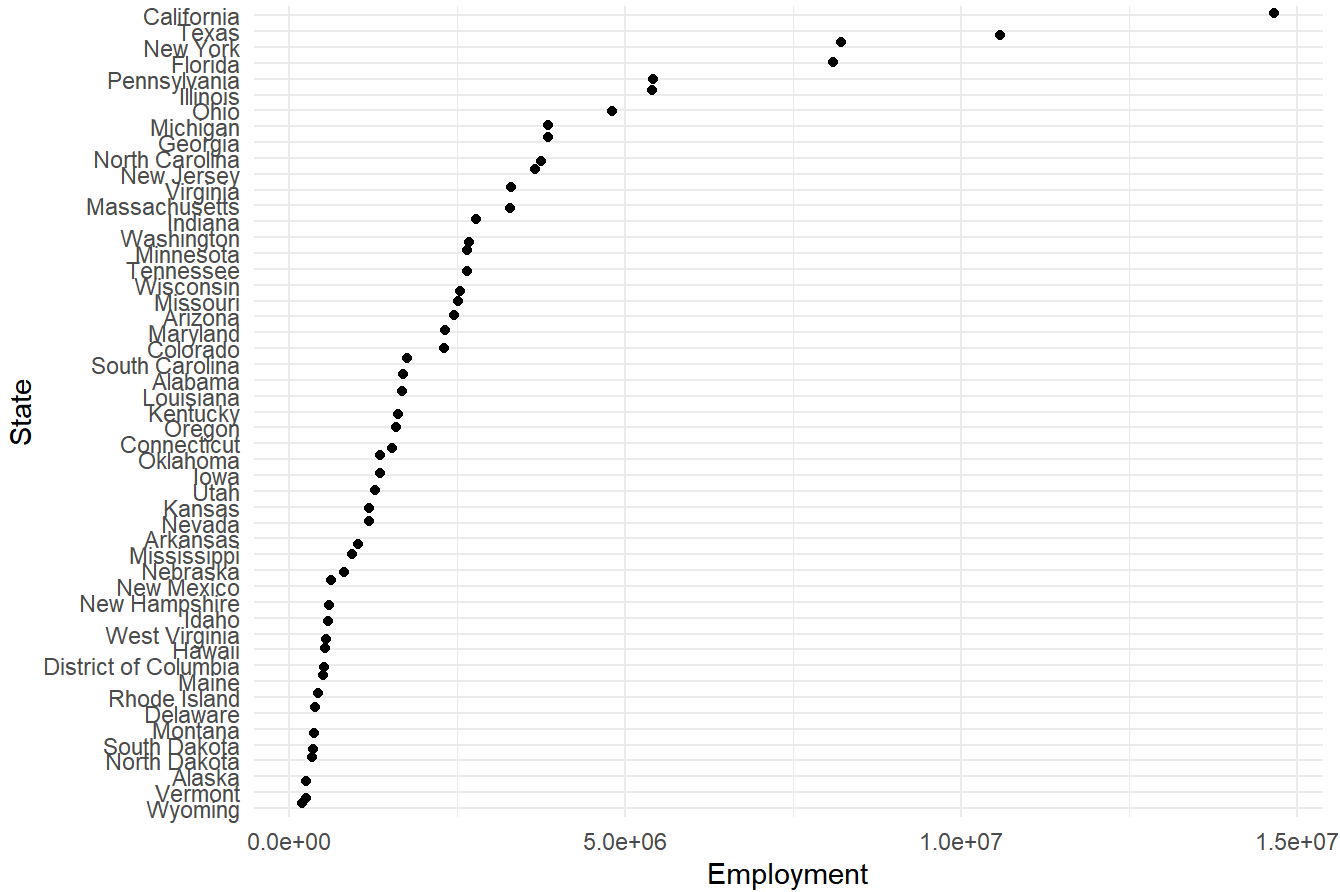
Employment by State in 2019



Employment by State in 2018



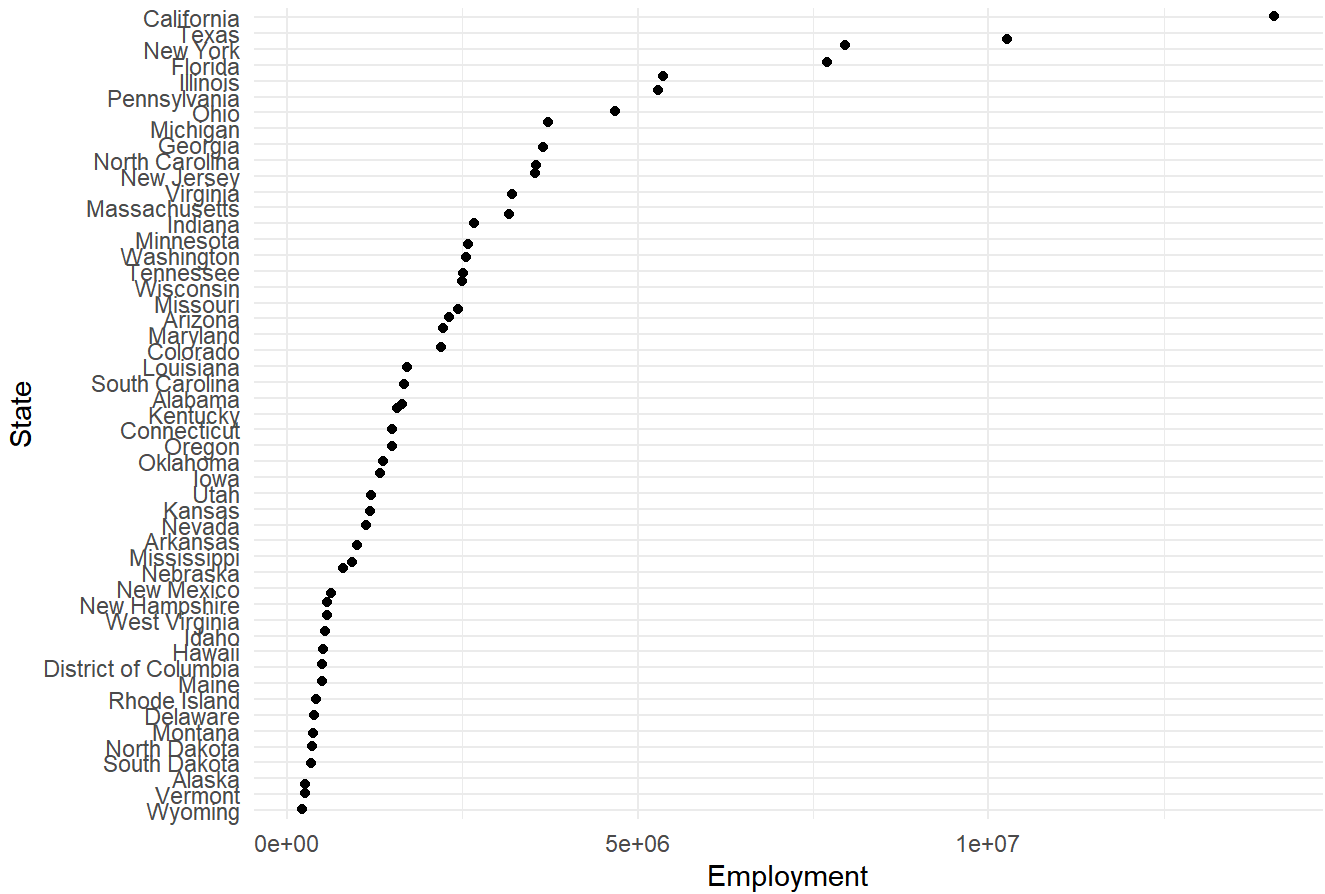
Employment by State in 2017



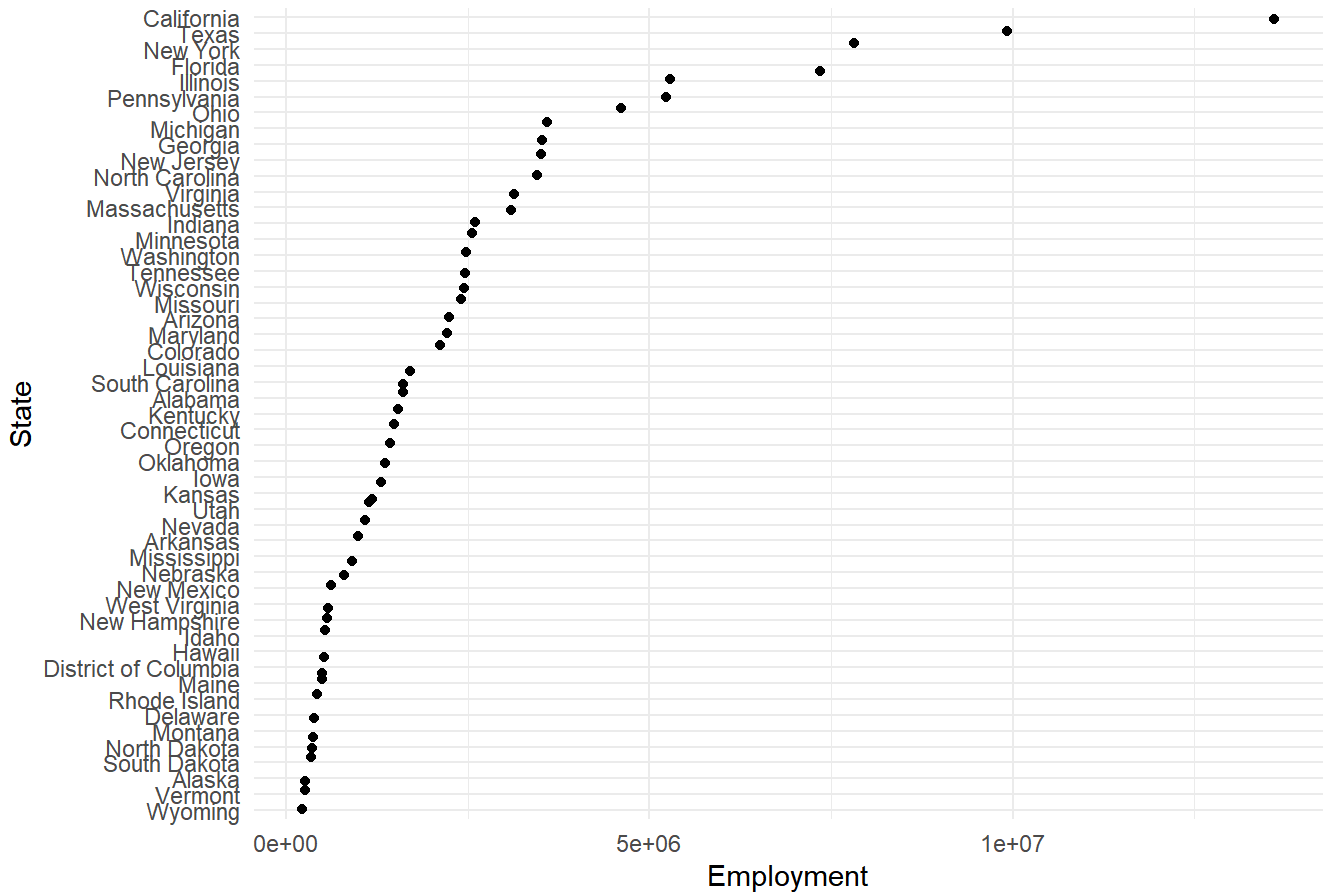


Scatter plot showing Employment (X-axis) versus State (Y-axis). The X-axis ranges from 0.0e+00 to 1.5e+07. The Y-axis lists 50 states and the District of Columbia, ordered by decreasing employment. The plot shows a strong positive correlation between the rank of the state (from highest to lowest employment) and its actual employment value.

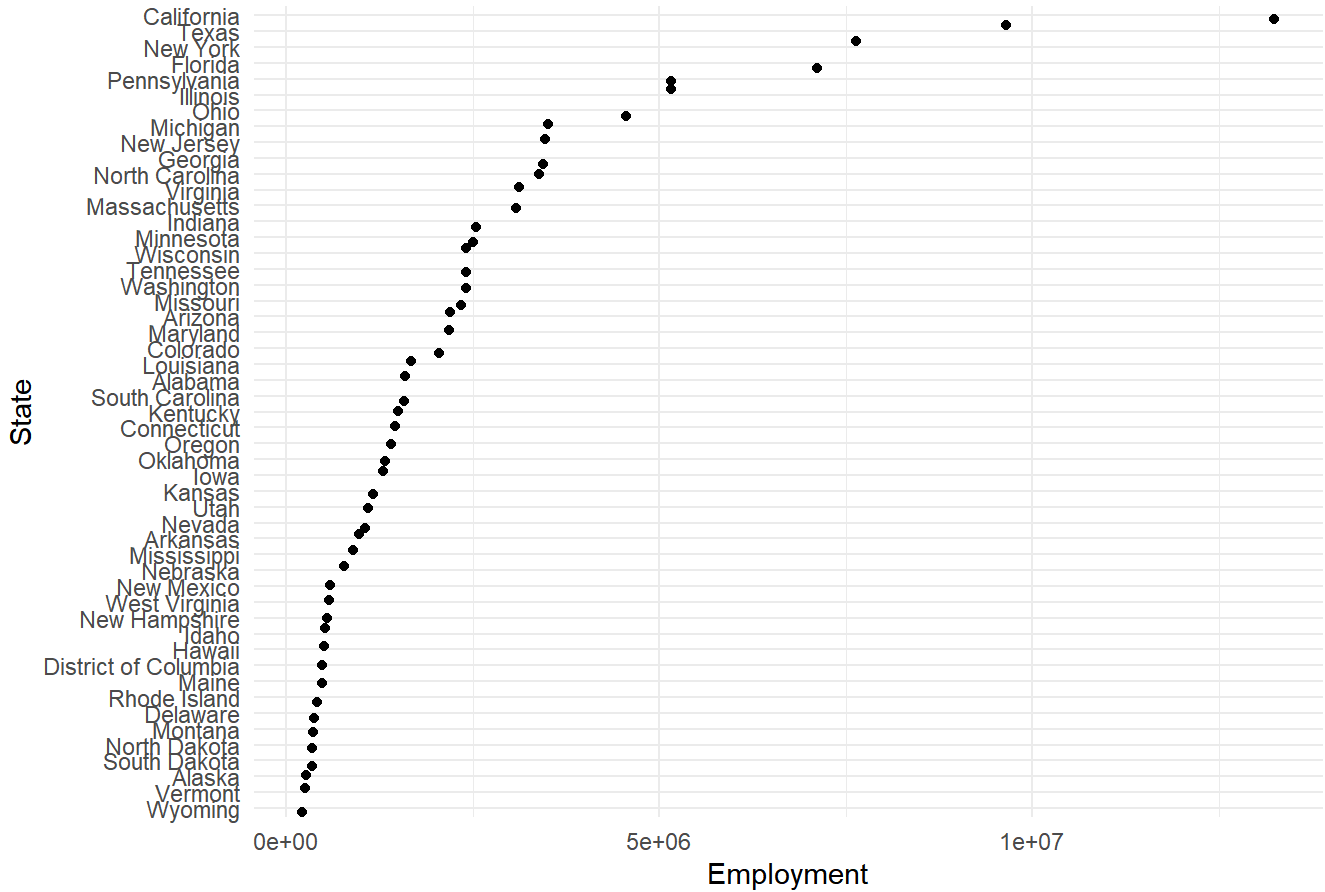
Employment by State in 2015



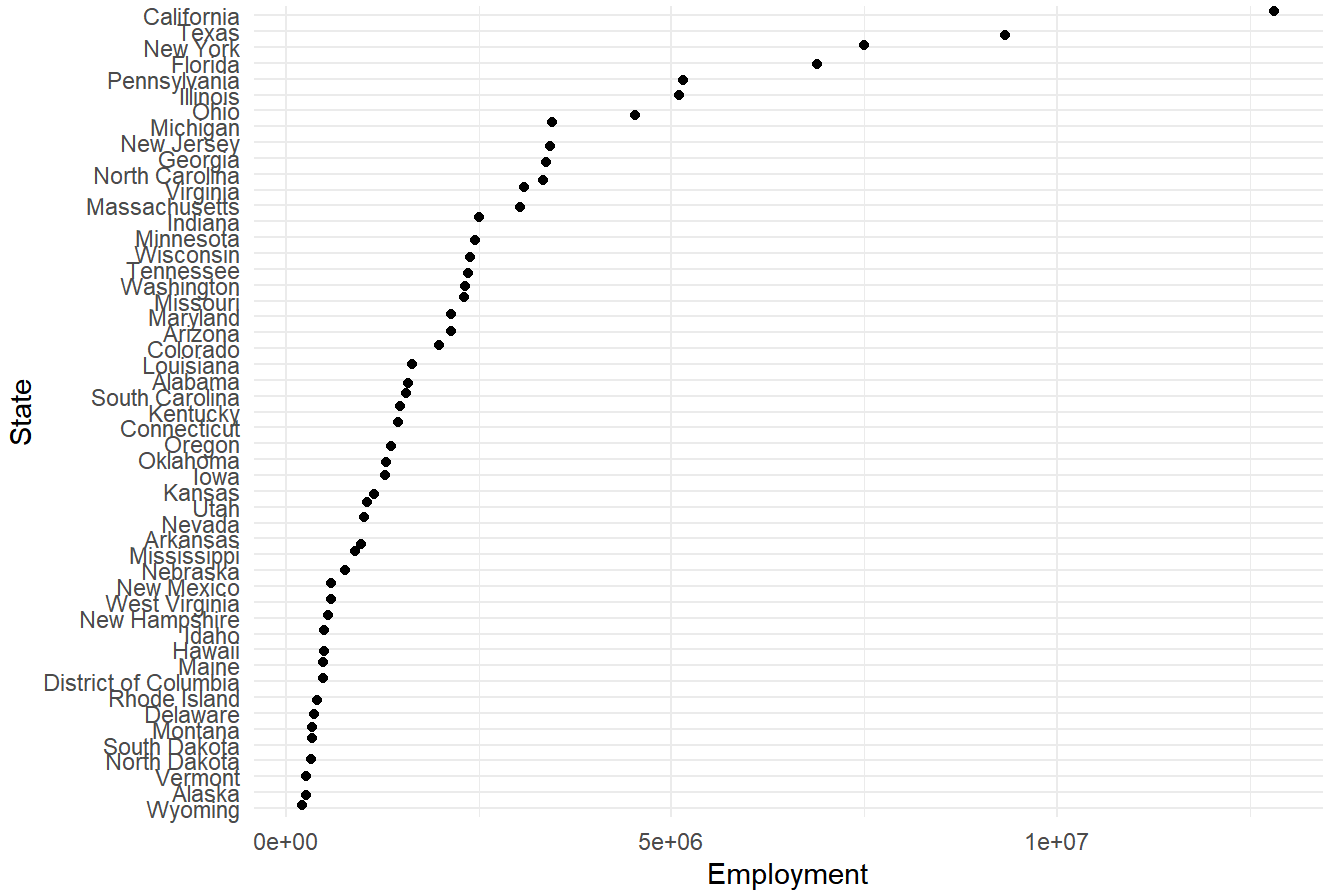
Employment by State in 2014



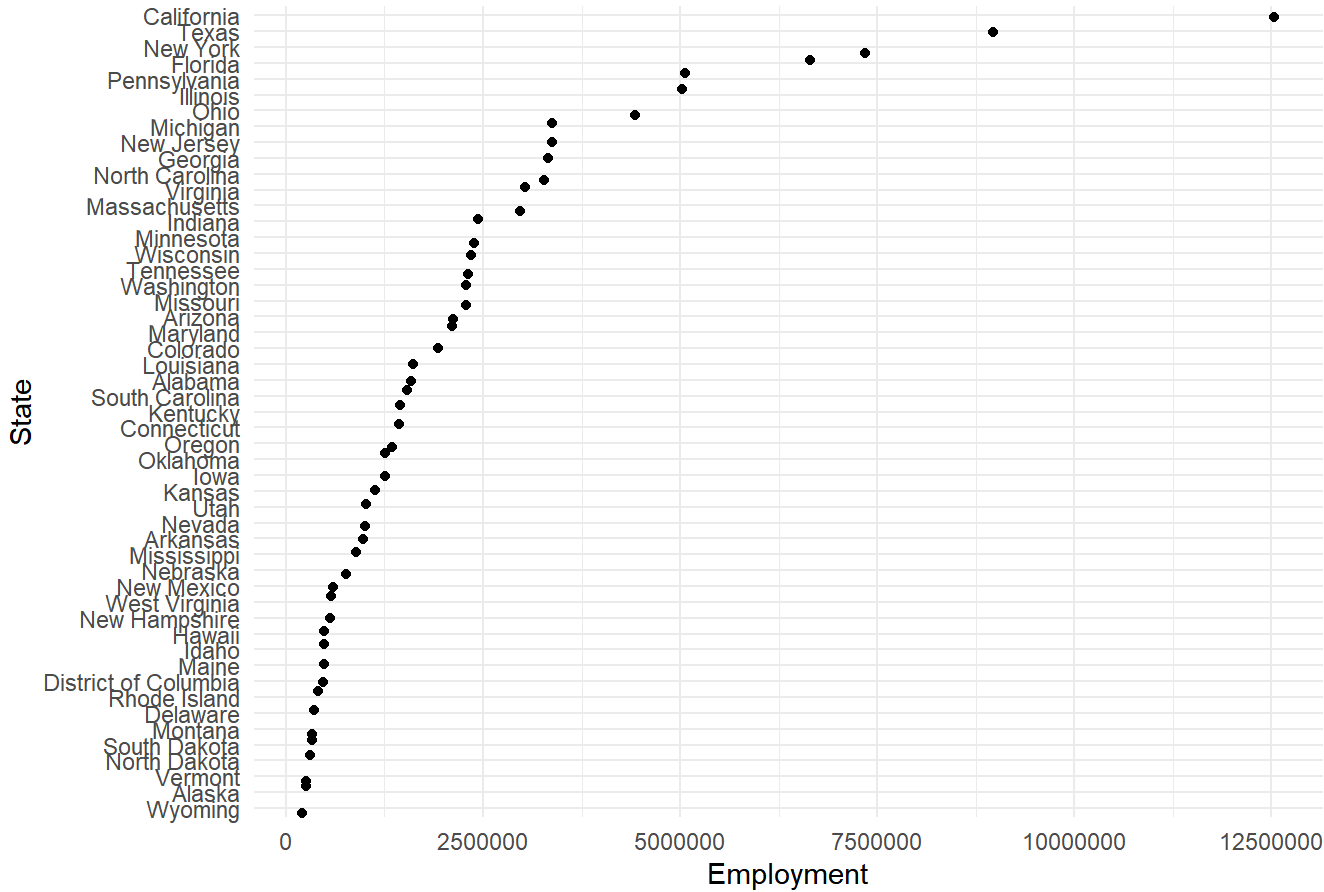
Employment by State in 2013



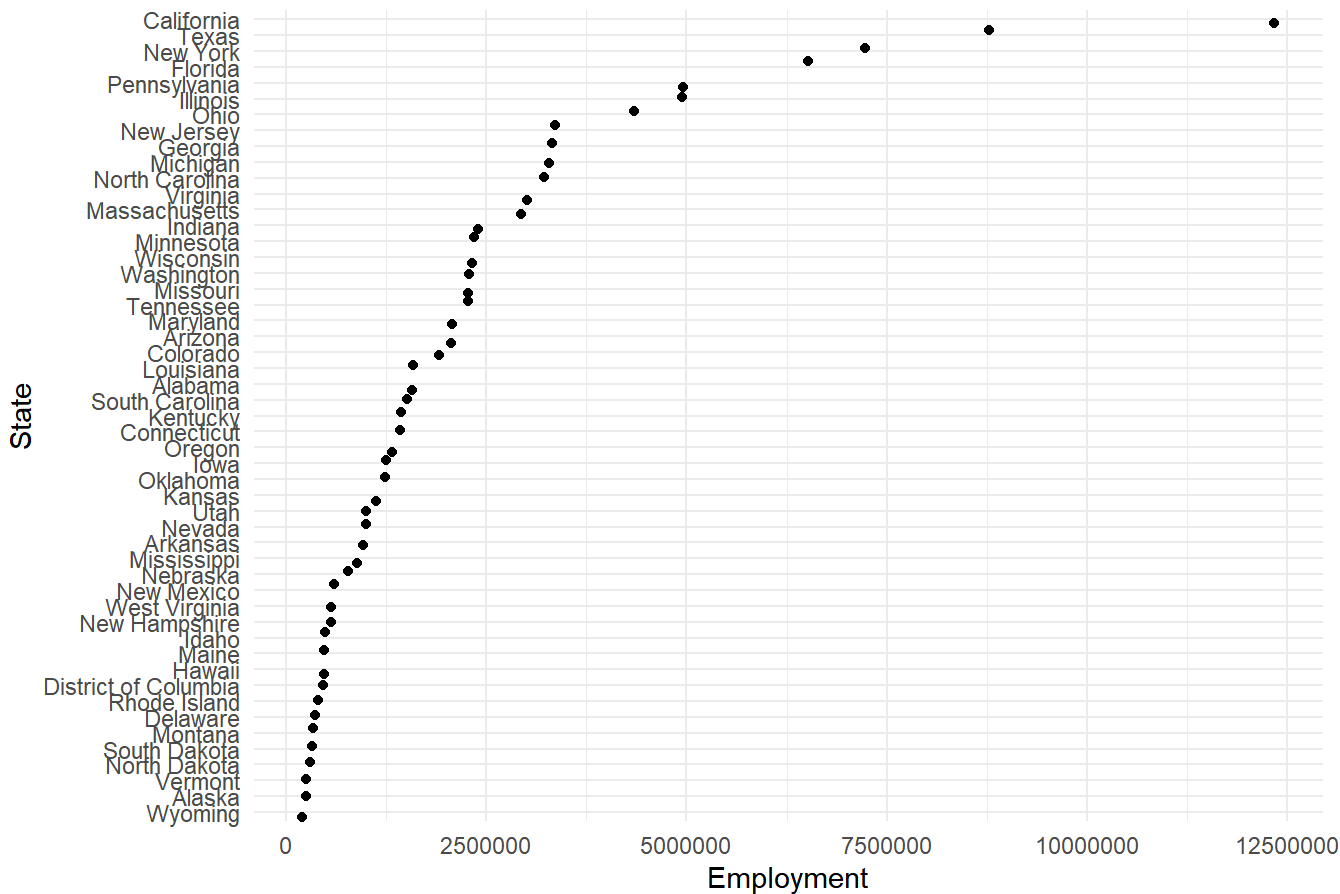
Employment by State in 2012



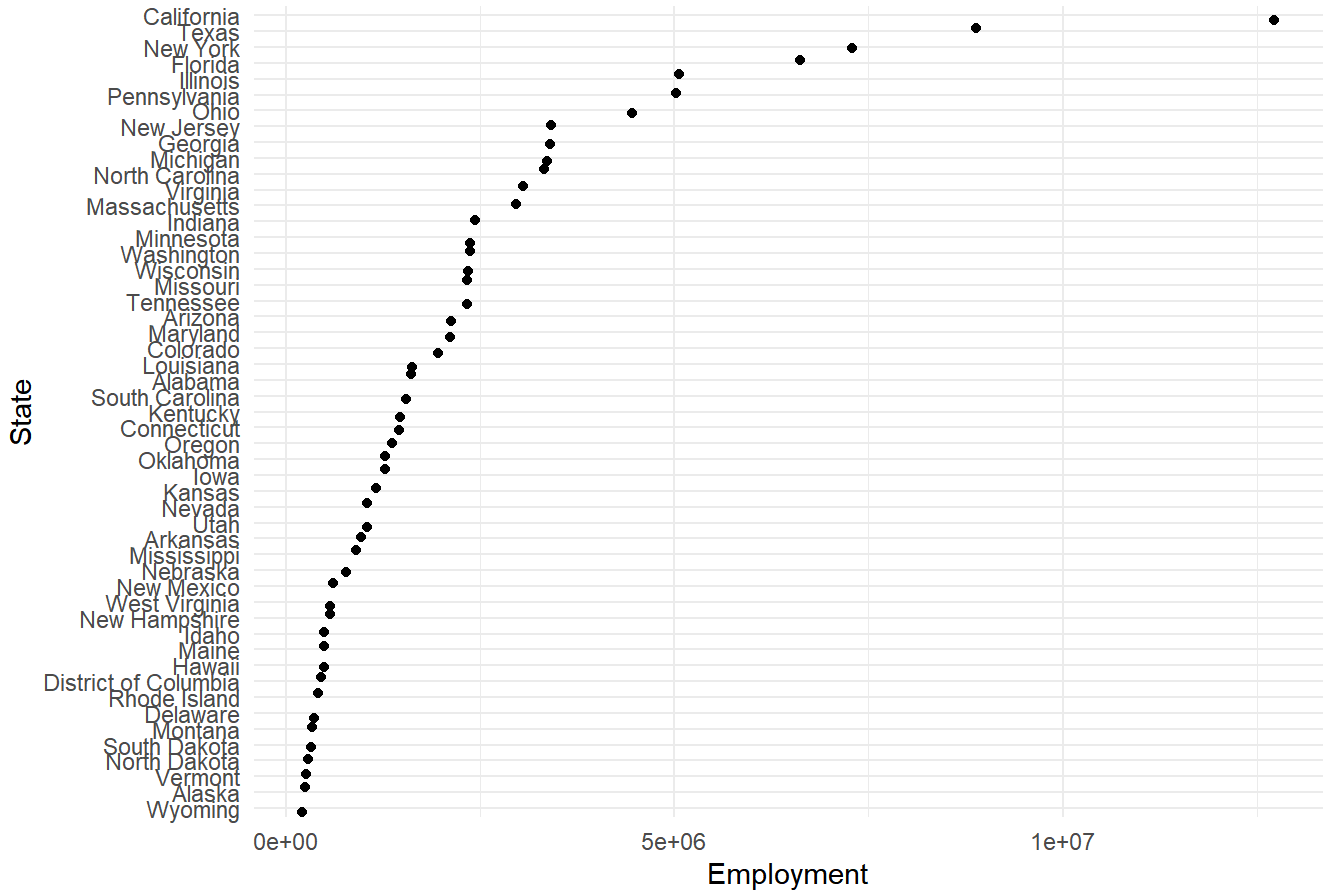
Employment by State in 2011



Employment by State in 2010

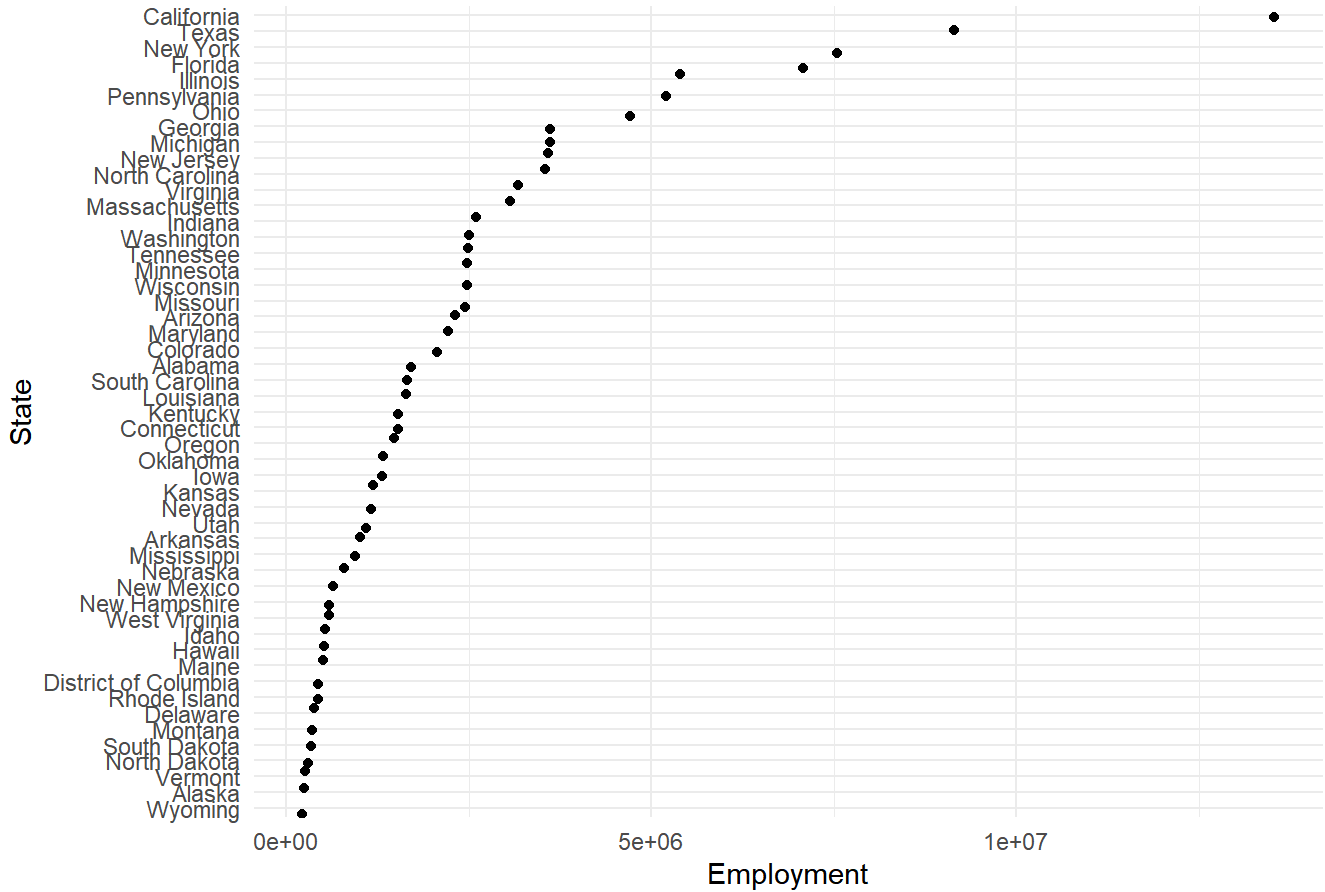


Employment by State in 2009

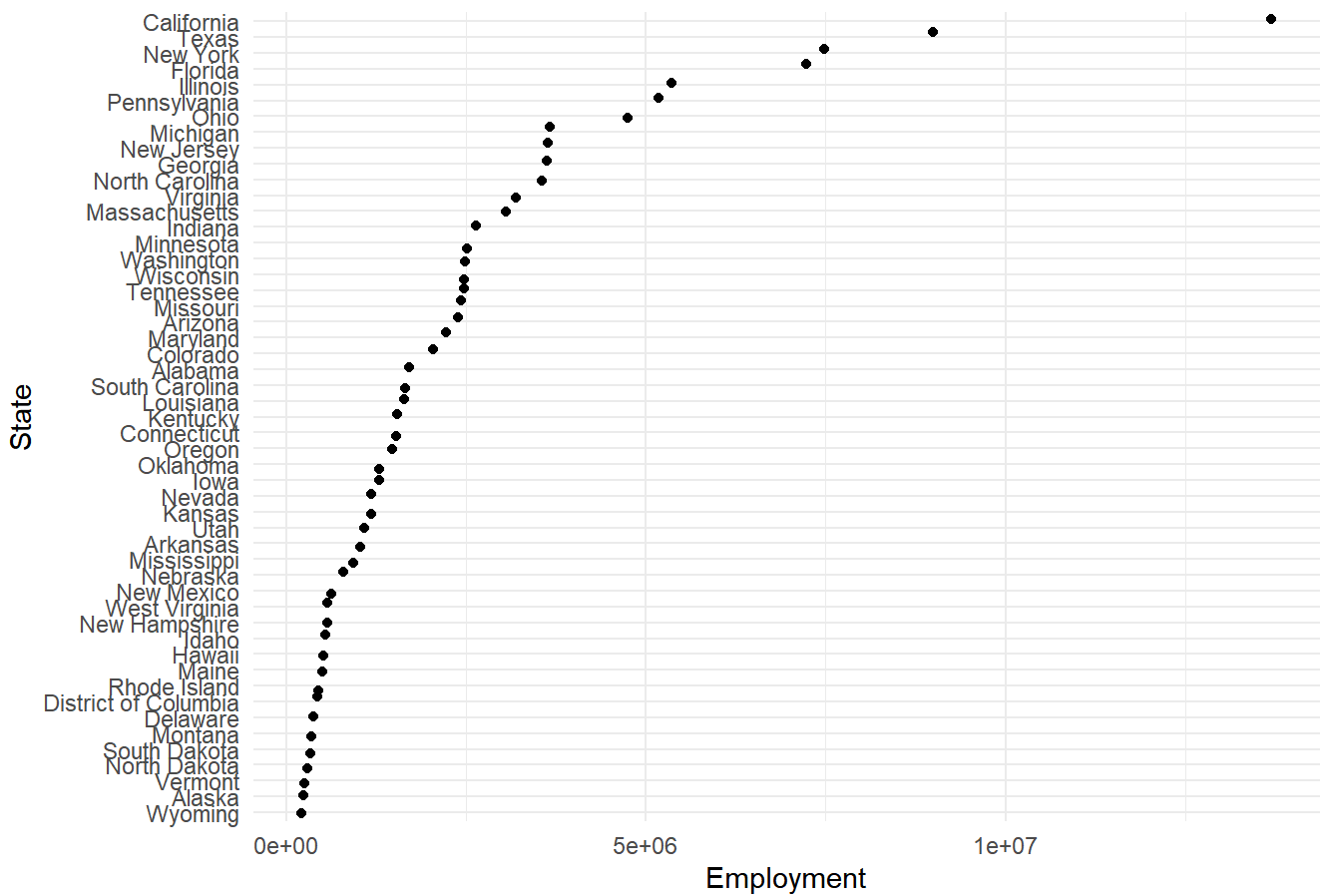




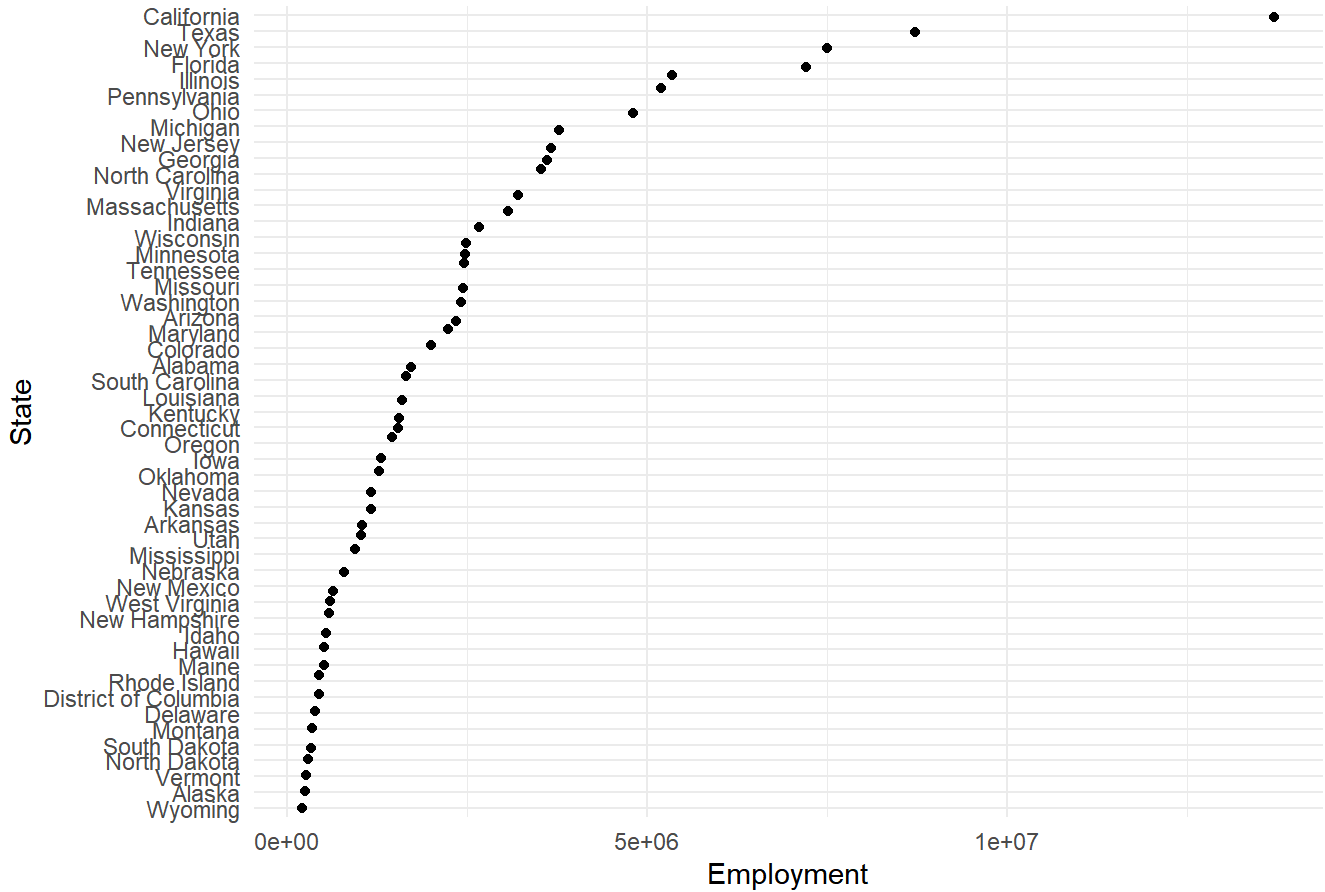
Employment by State in 2008



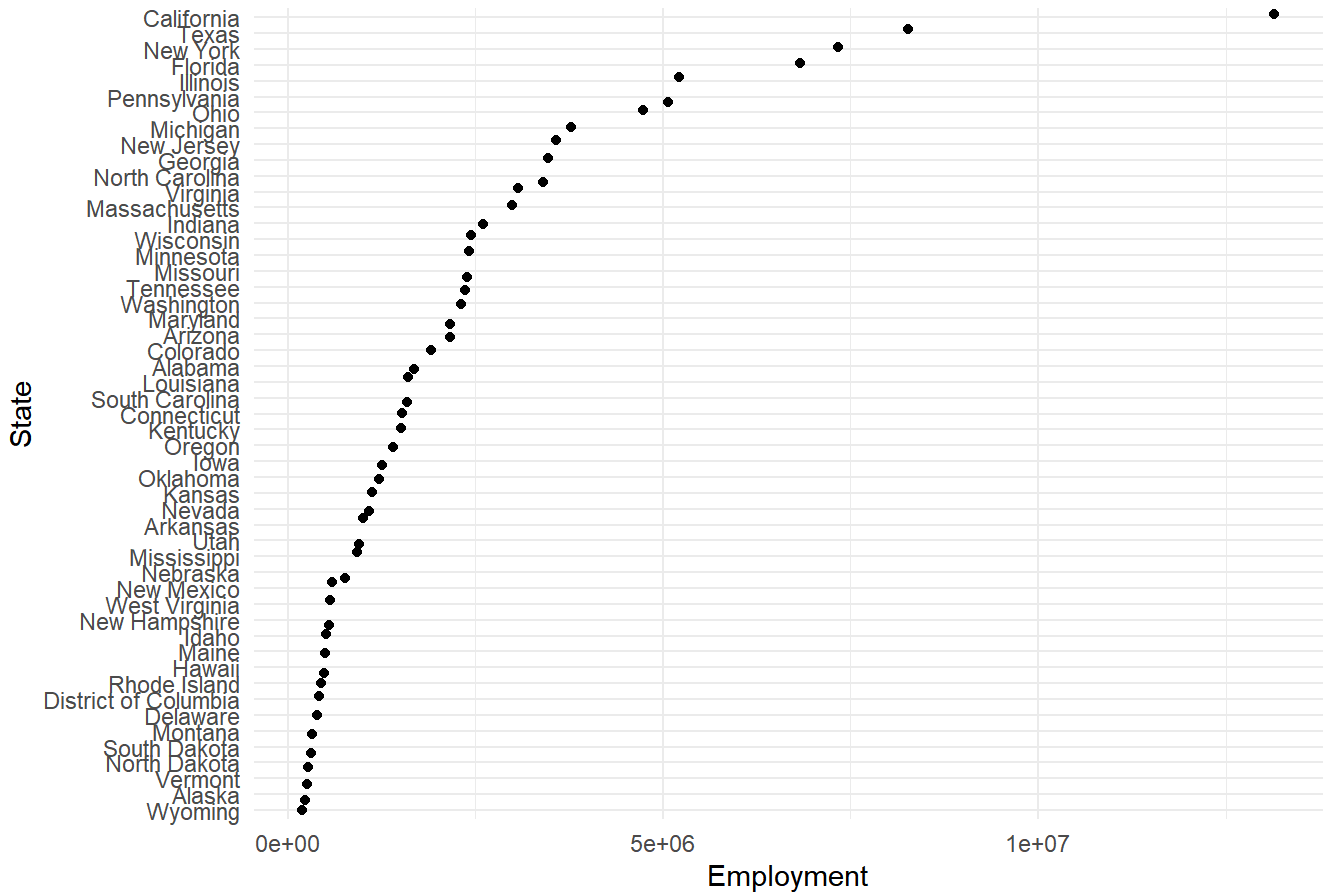
Employment by State in 2007



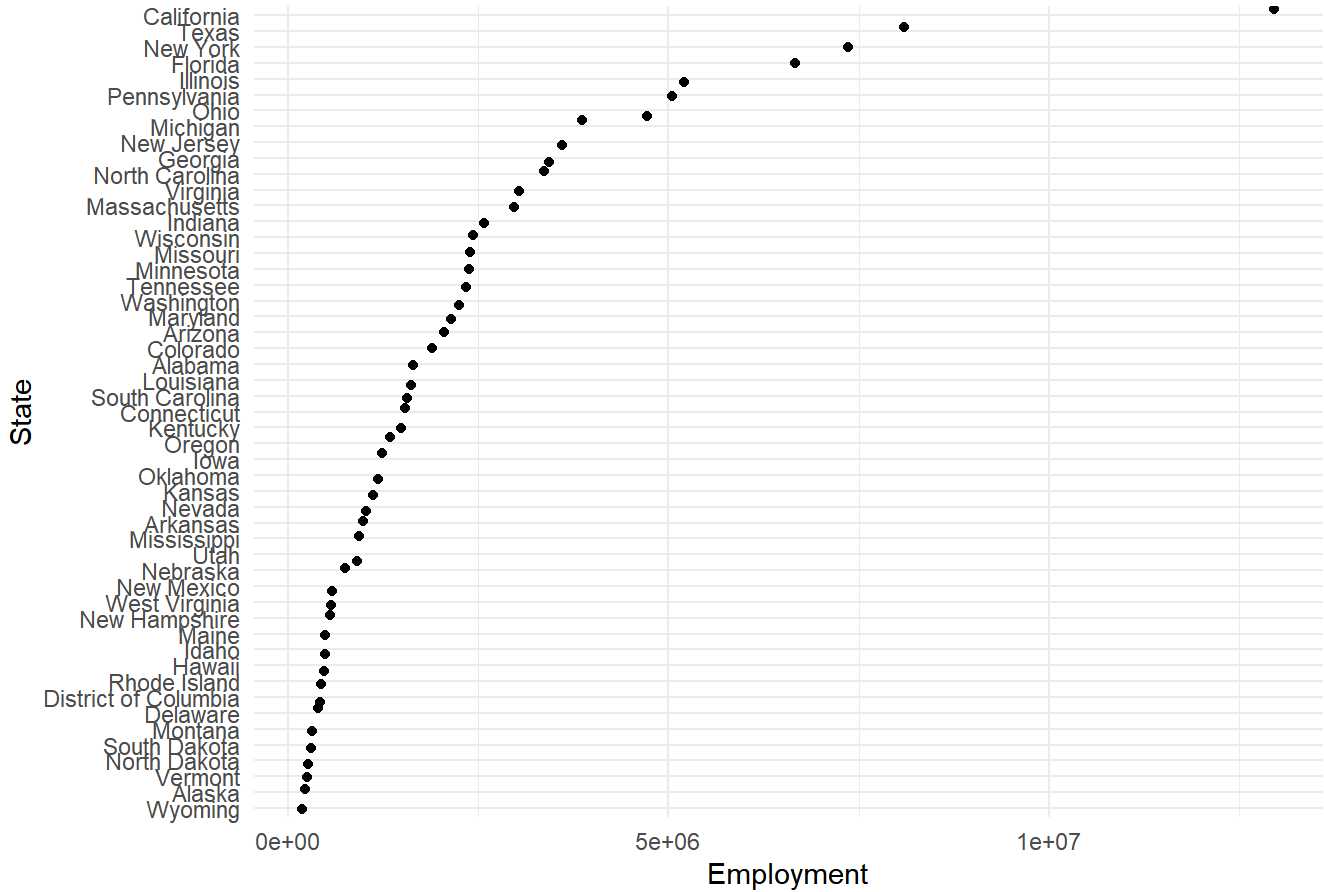
Employment by State in 2006



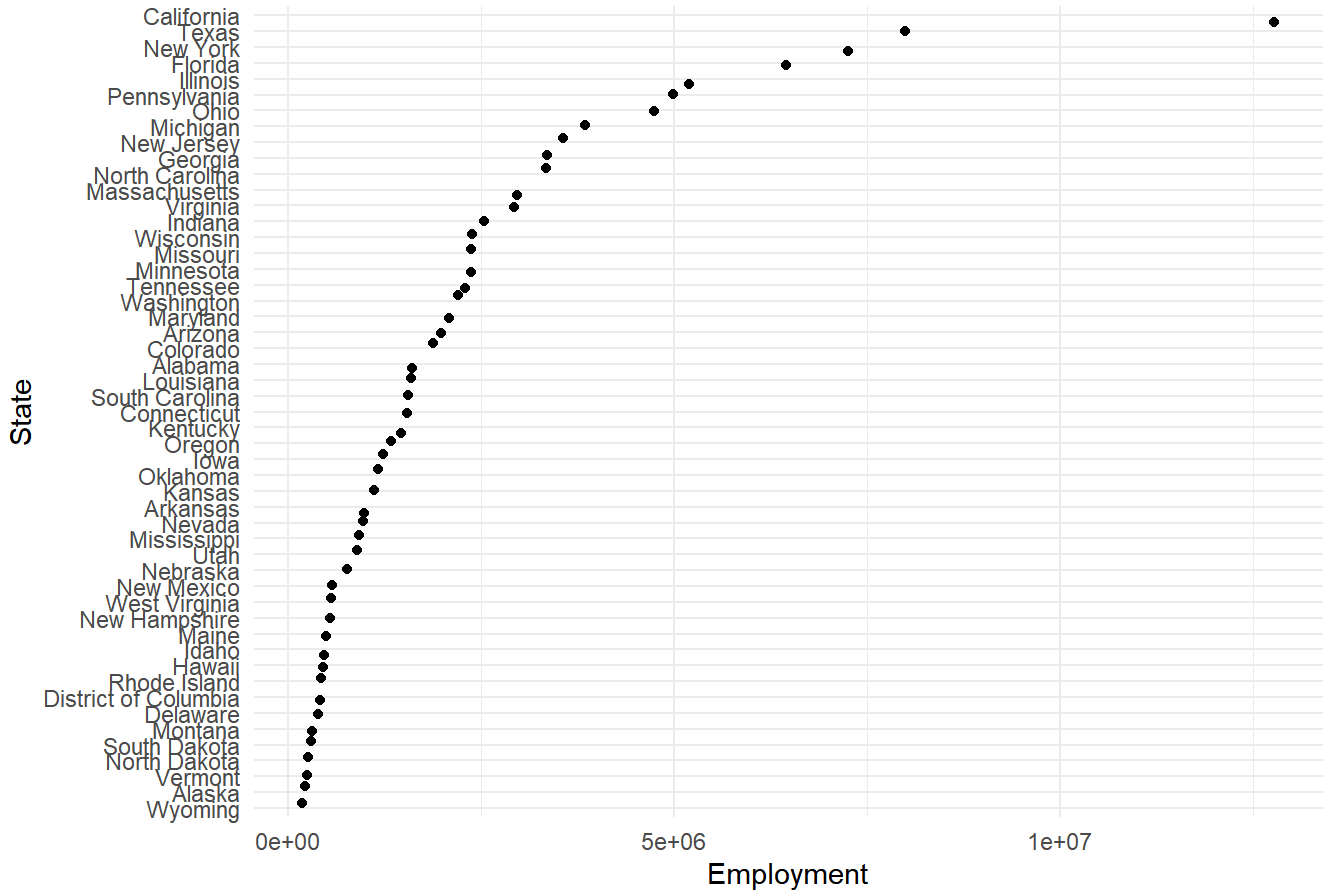
Employment by State in 2005



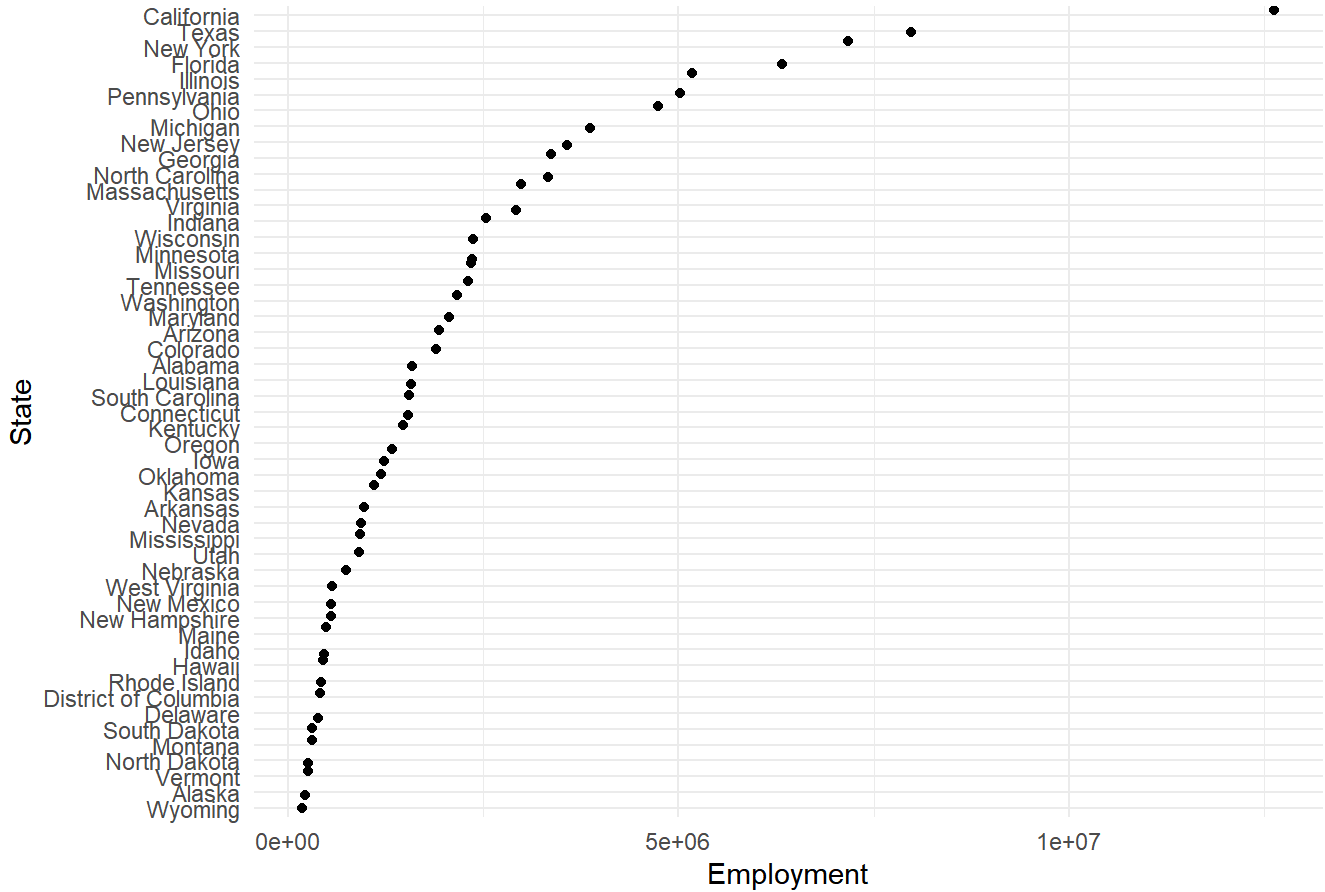
Employment by State in 2004



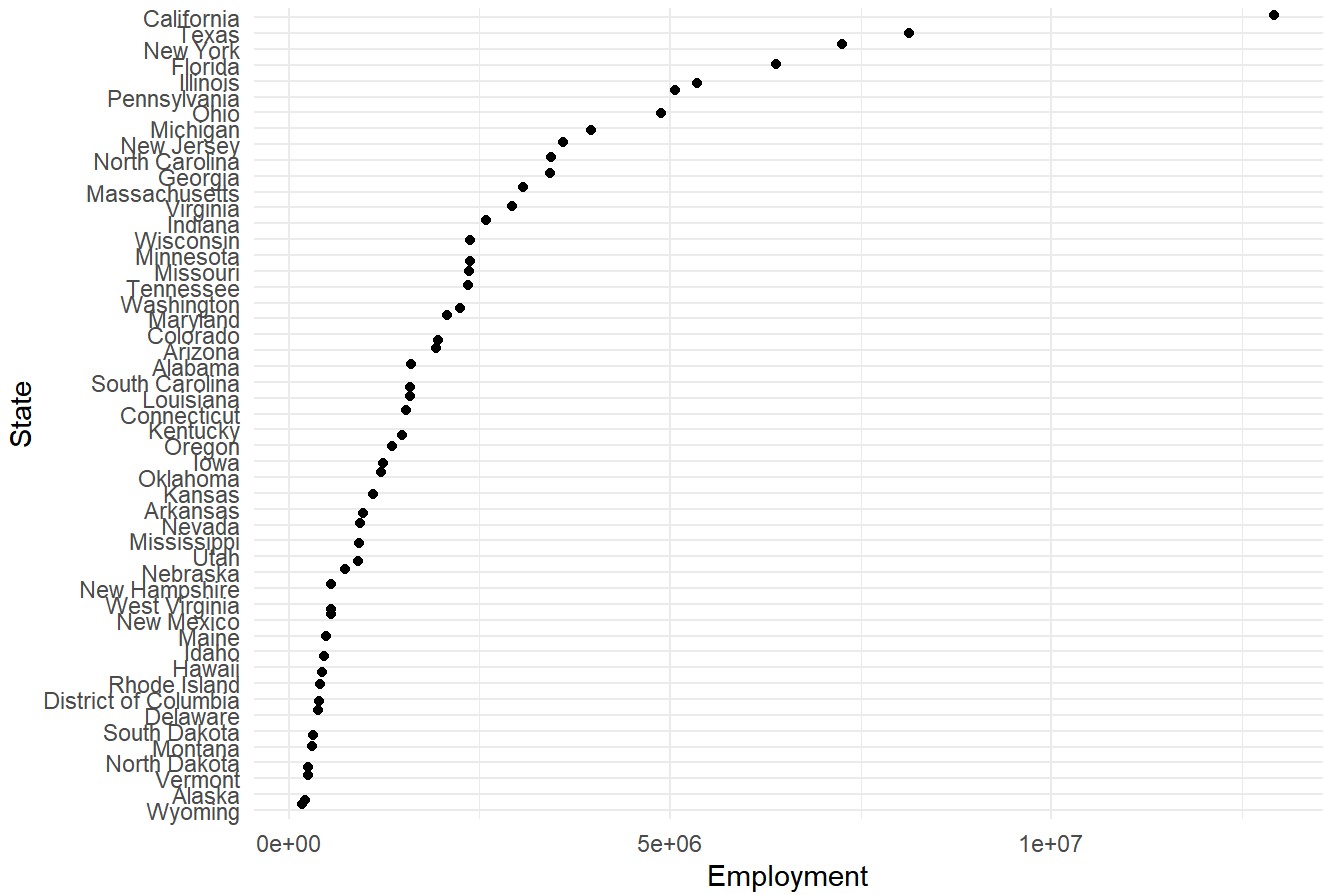
Employment by State in 2003



Employment by State in 2002

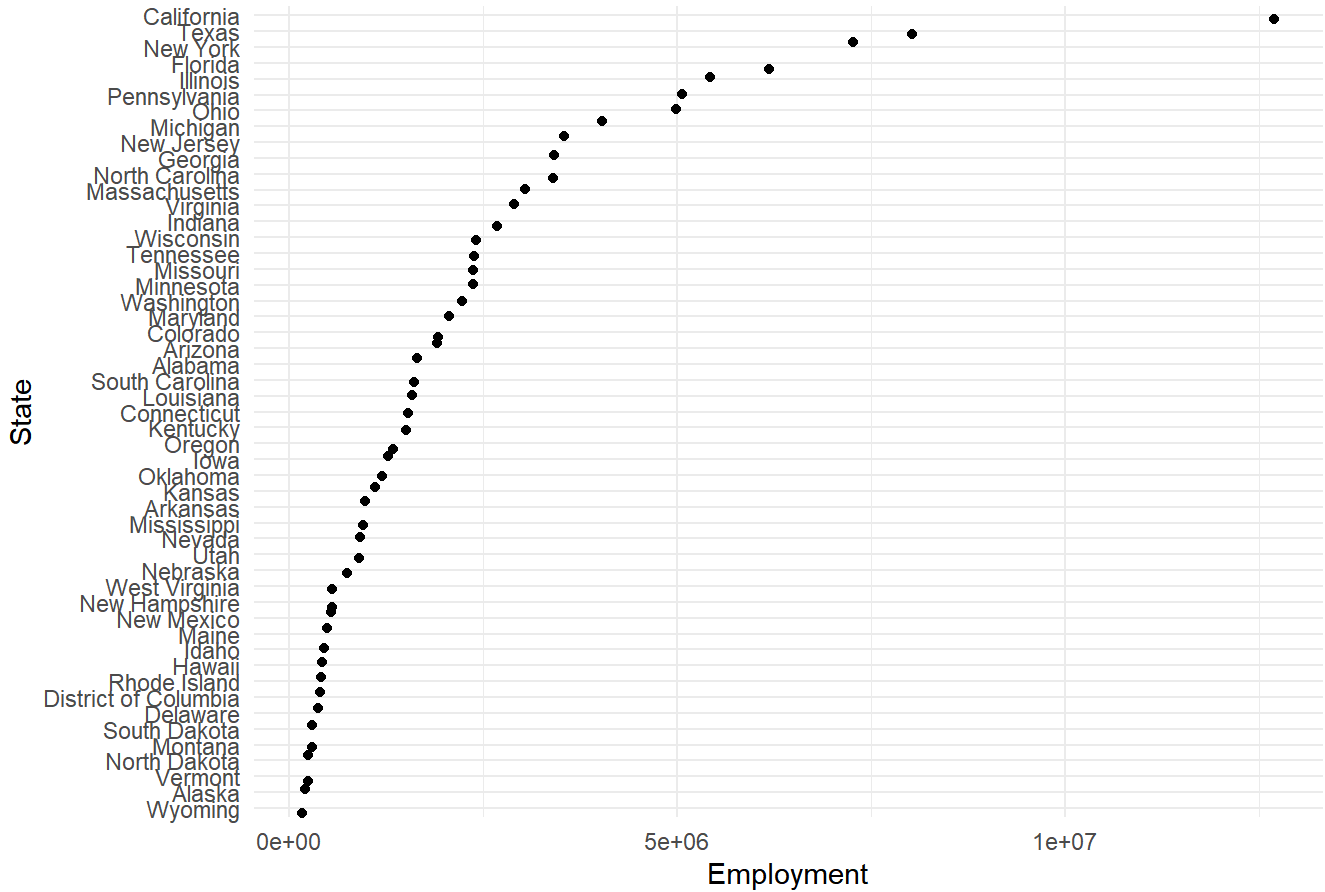


Employment by State in 2001

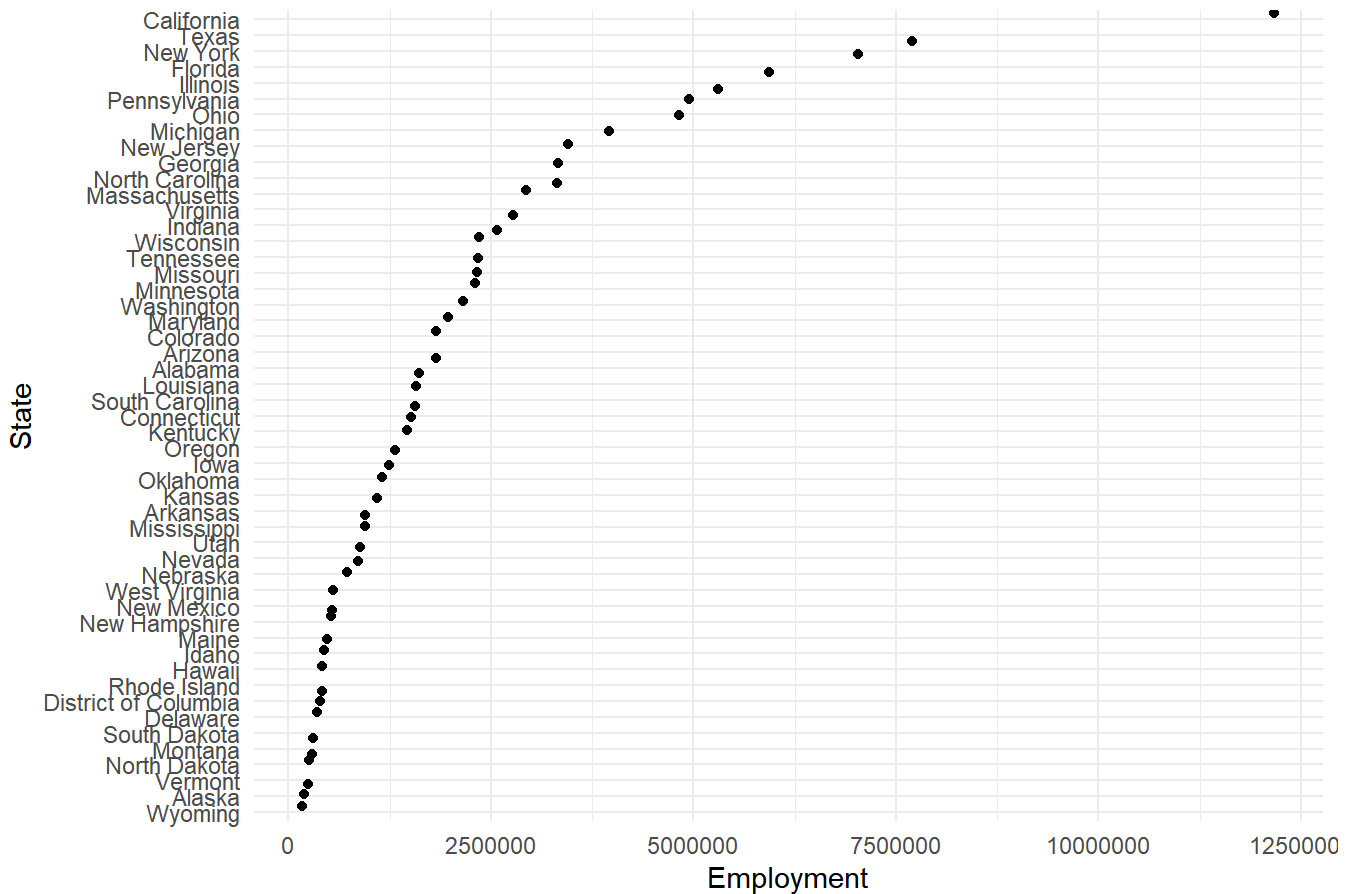




Employment by State in 2000



## Employment by State in 1999



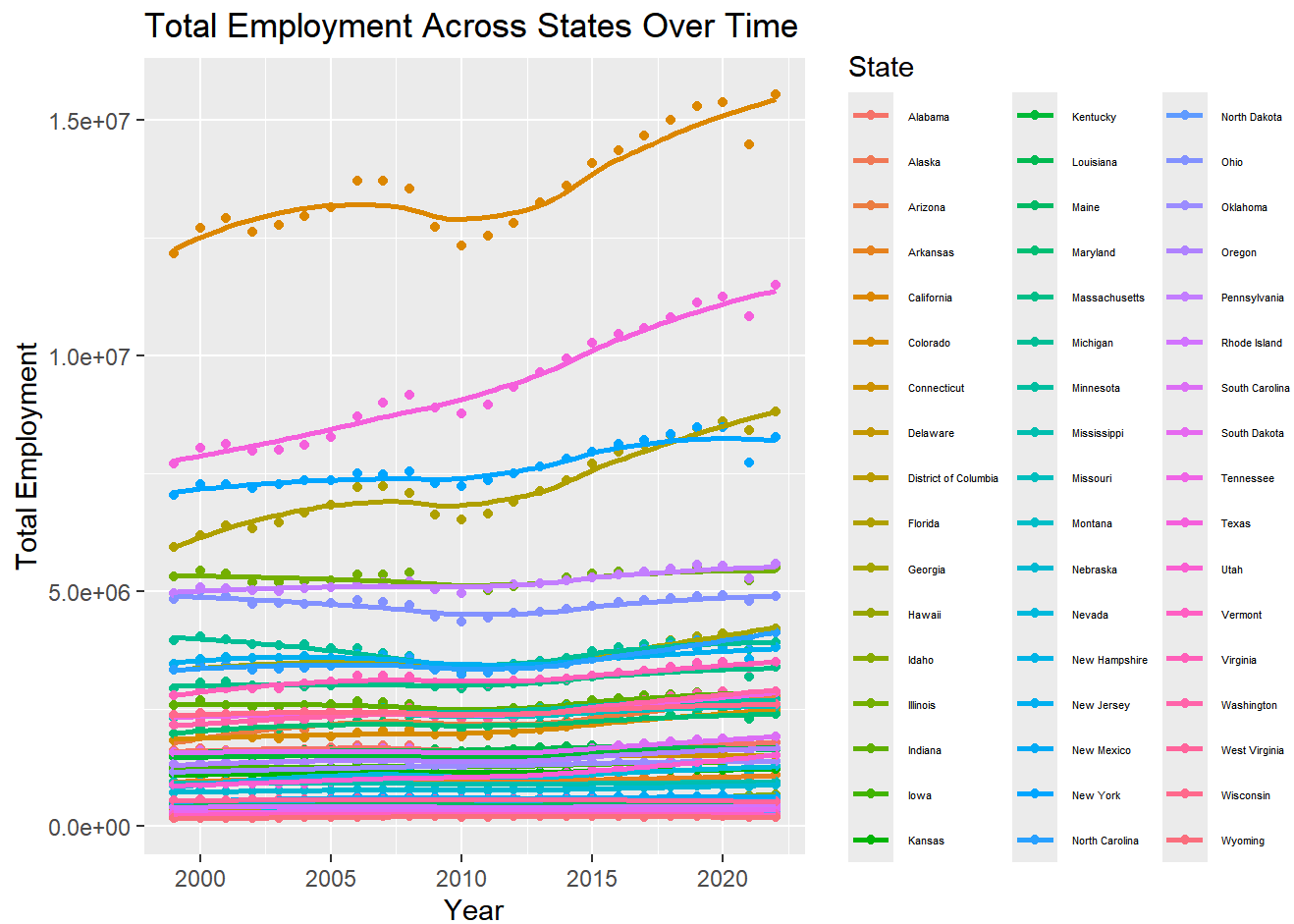
## Separate plots for each variable trend over time, with separate lines for every state

```
variables <- c("Total Employment", "Information Sector", "Net Job Creation",
              "Establishment Births", "Establishment, Firm Deaths",
              "Number of Utility Patents", "Proportion of S&E",
              "Number of Firms Receiving VC")

# Loop through each variable and generate the plot
suppressWarnings(for (v in variables) {
  plot <- ggplot(full, aes(x = Year, y = .data[[v]], color = State)) +
    geom_point() +
    geom_smooth(se = FALSE) +
    labs(title = paste(v, "Across States Over Time"),
         x = "Year",
         y = v) +
    theme(legend.position = "right",
          legend.text = element_text(size = 4))

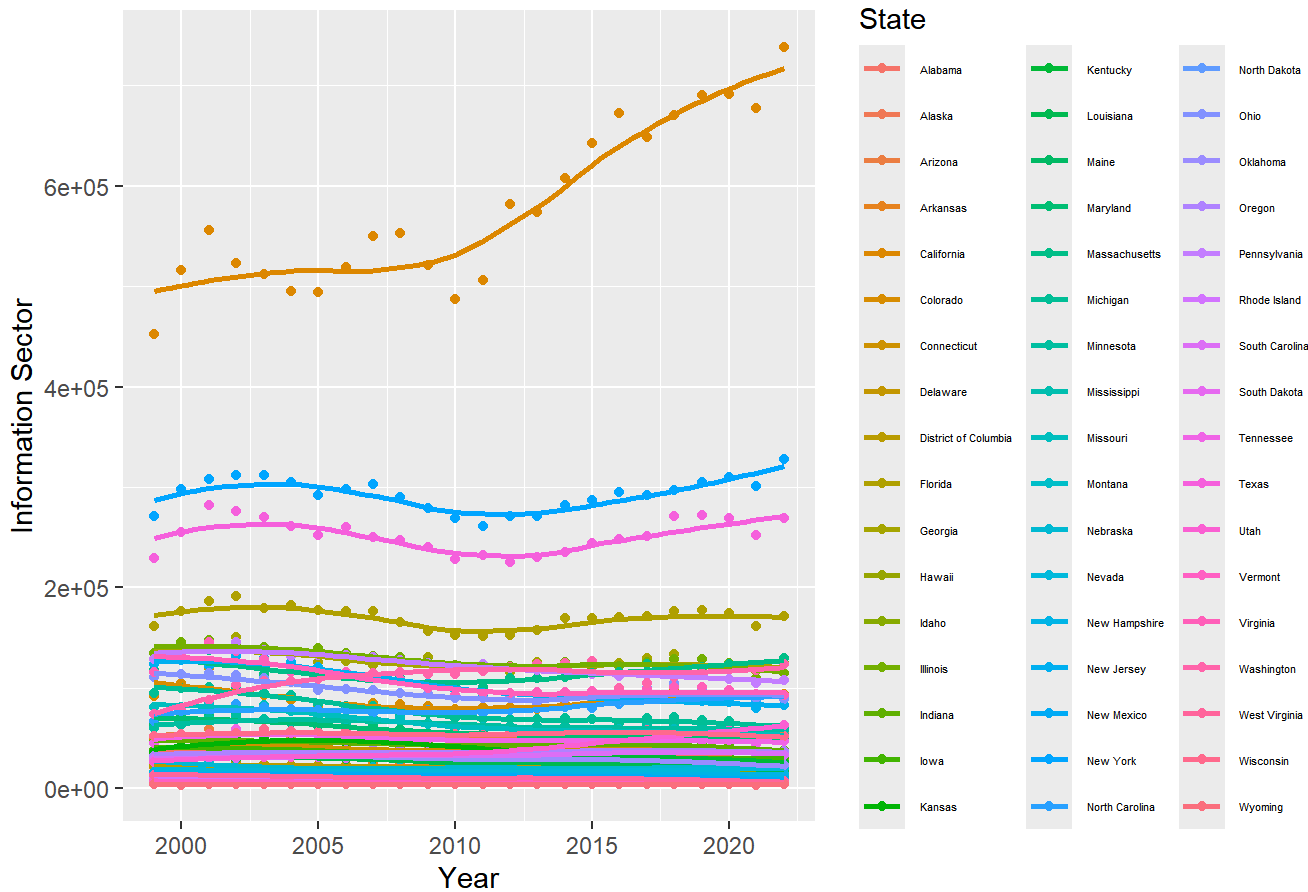
  print(plot)
})
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



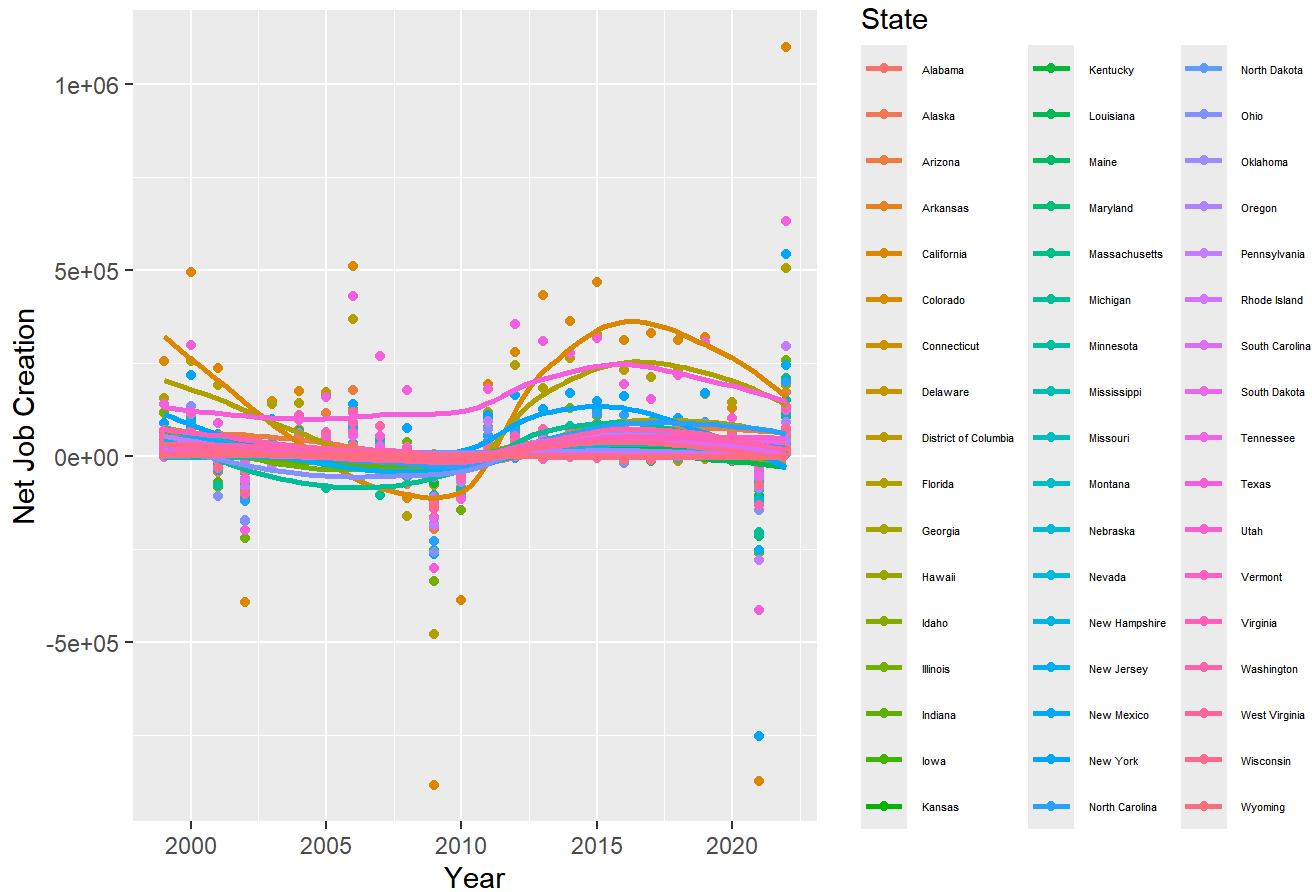
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

## Information Sector Across States Over Time

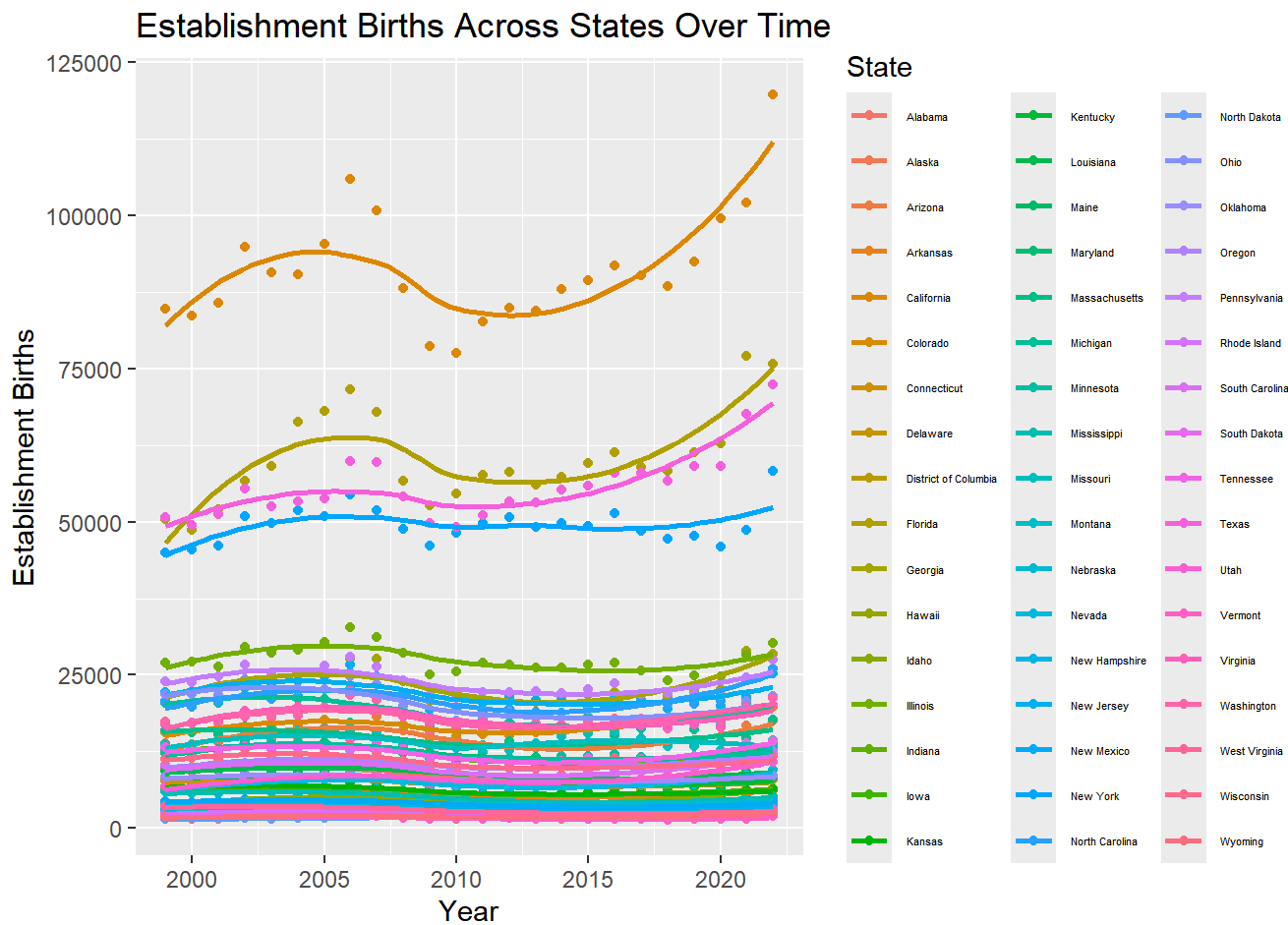


``geom_smooth()`` using method = 'loess' and formula = 'y ~ x'

## Net Job Creation Across States Over Time

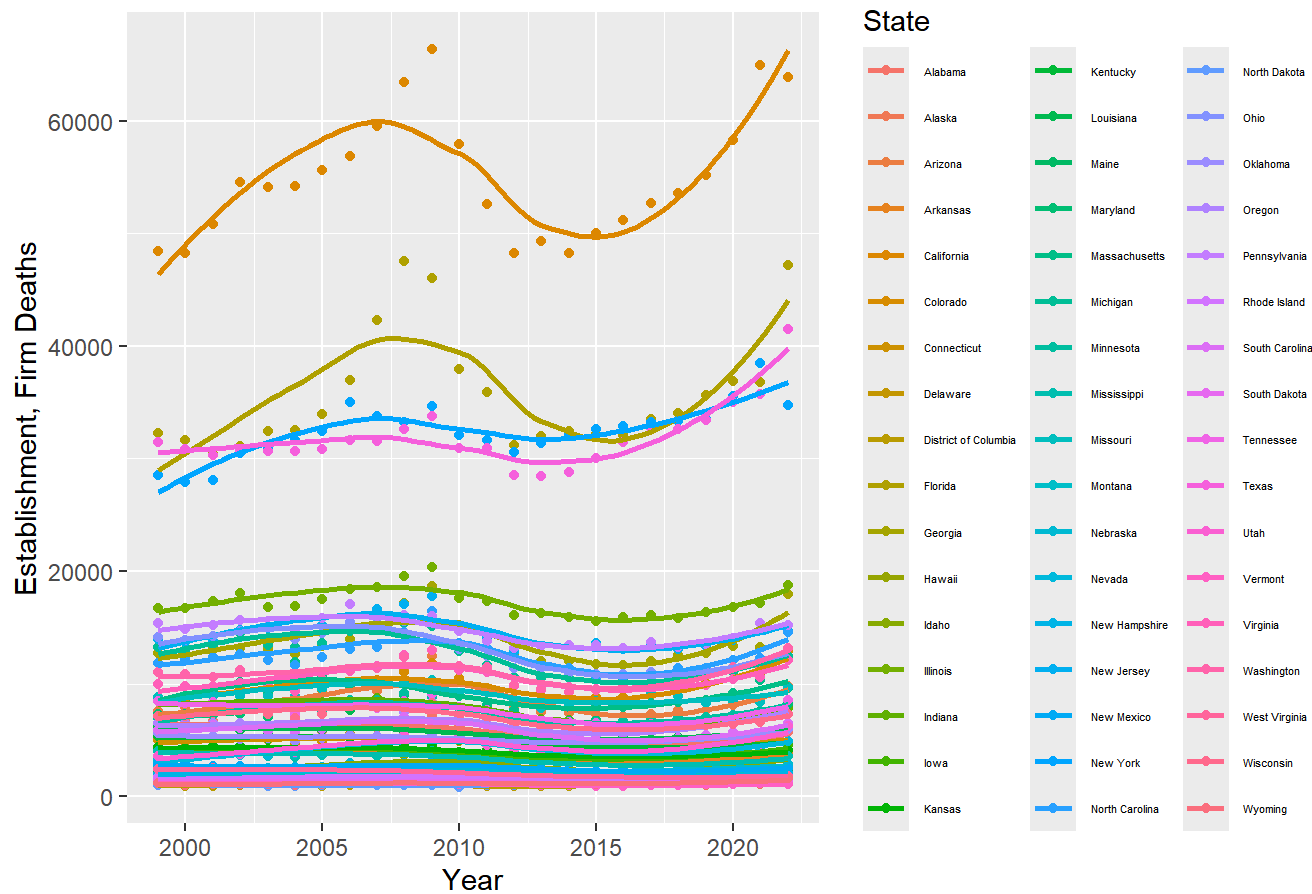


`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



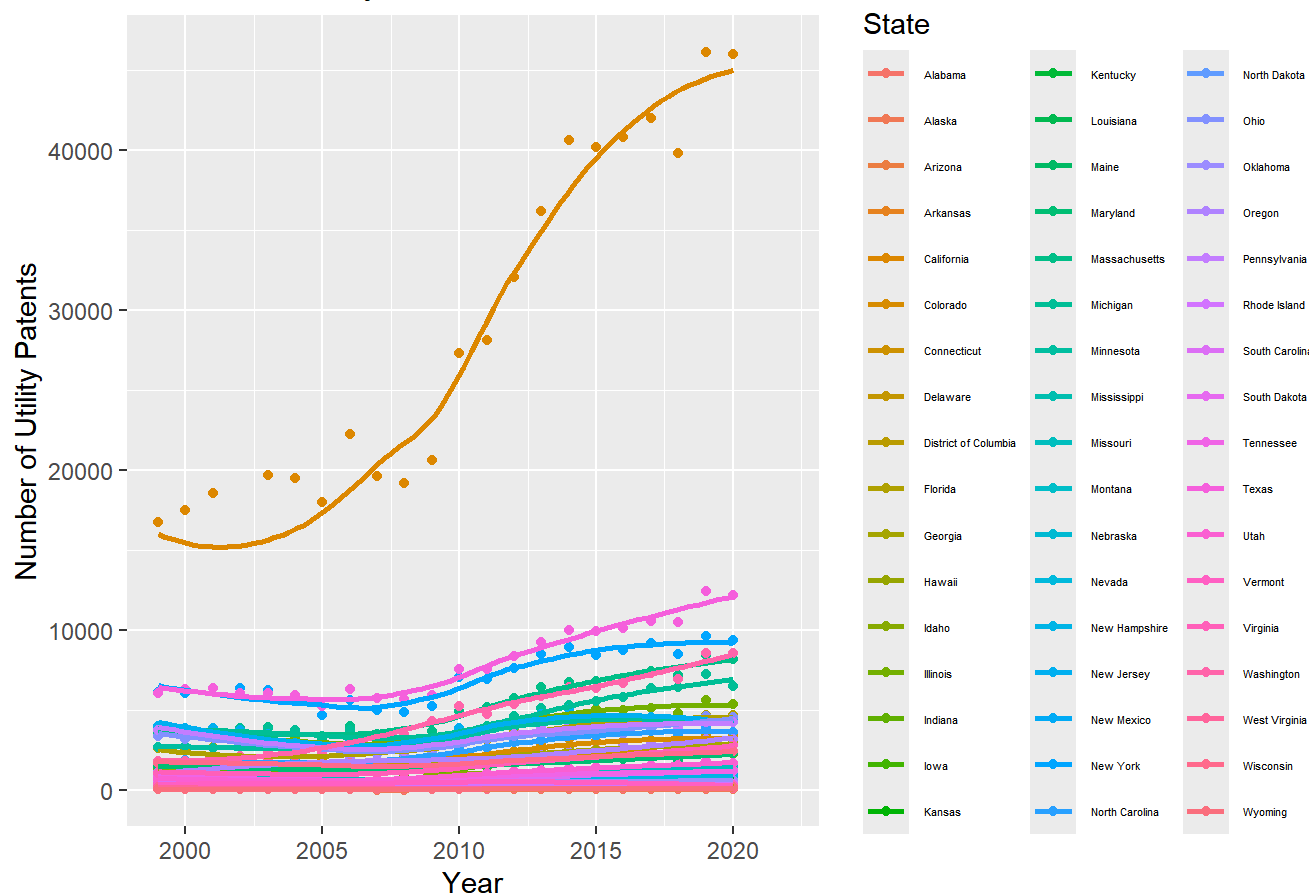
``geom_smooth()`` using method = 'loess' and formula = 'y ~ x'

# Establishment, Firm Deaths Across States Over Time



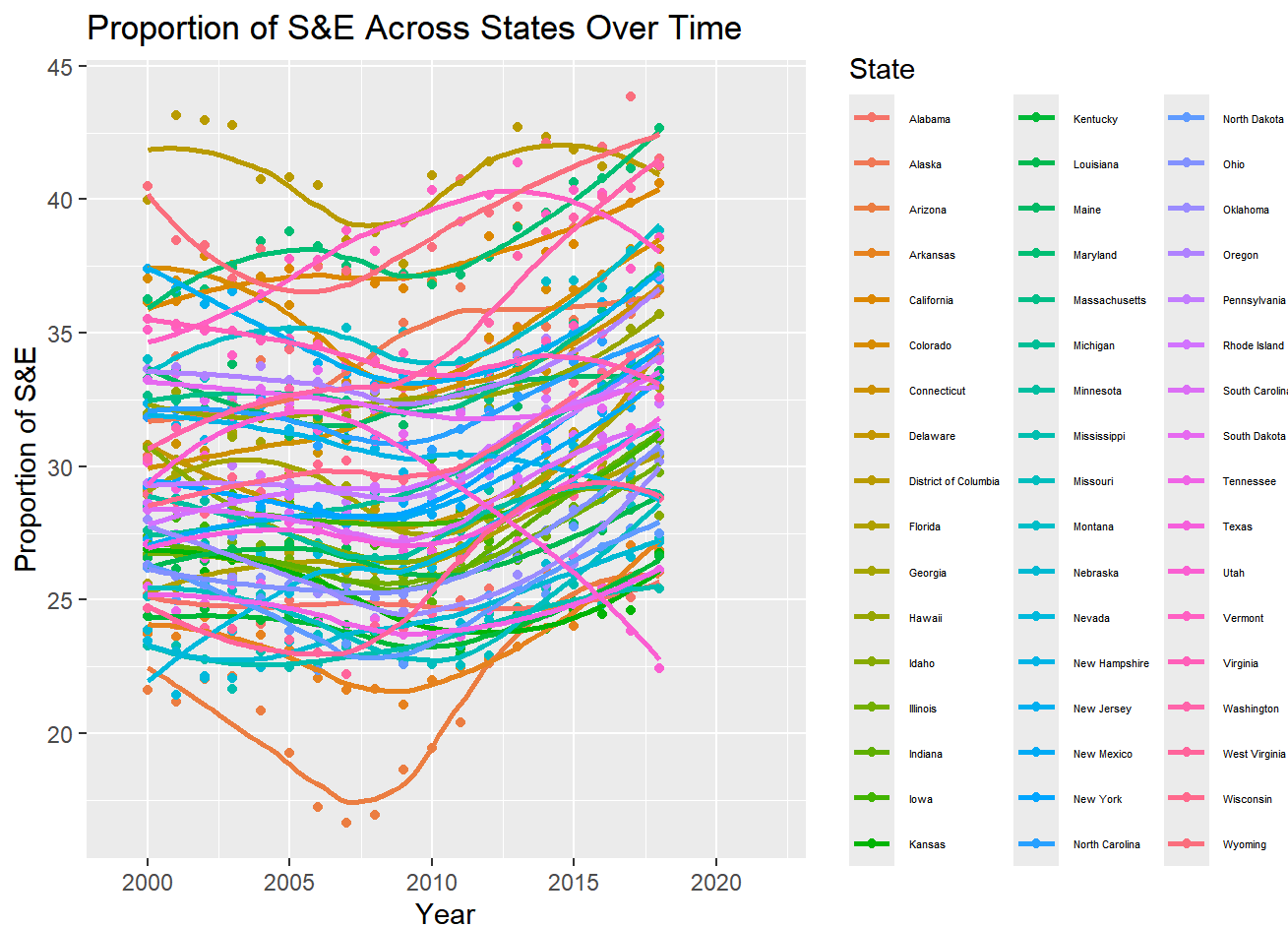
``geom_smooth()`` using method = 'loess' and formula = 'y ~ x'

## Number of Utility Patents Across States Over Time



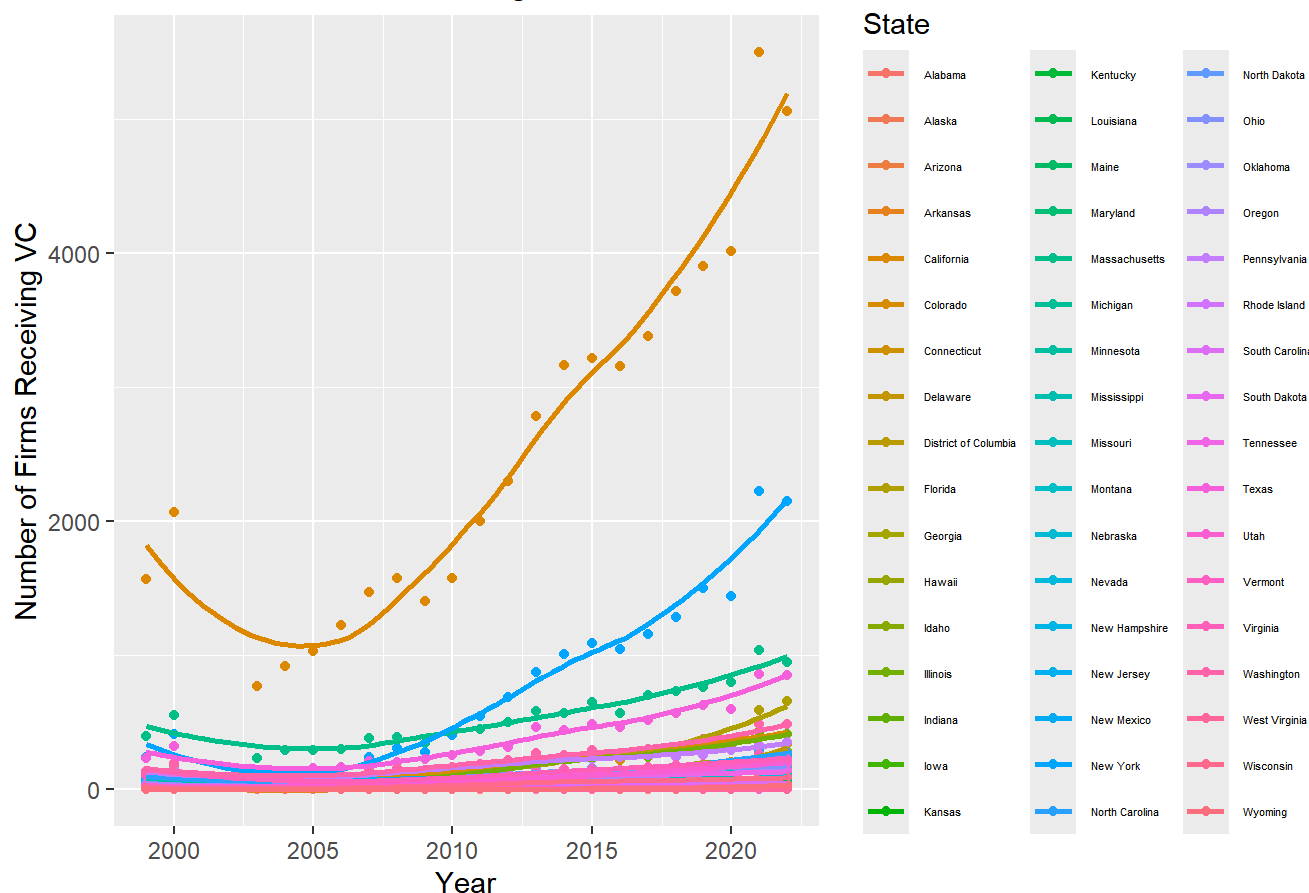
``geom_smooth()`` using method = 'loess' and formula = 'y ~ x'





``geom_smooth()`` using method = 'loess' and formula = 'y ~ x'

## Number of Firms Receiving VC Across States Over Time



## Separate plots for each variable (summarized over time) across states

```
#summary plot

full_summary <- full |>
  group_by(State) |>
  summarise(employment = mean(`Total Employment`, na.rm = TRUE),
            sector_employment = mean(`Information Sector`, na.rm = TRUE),
            jobs = mean(`Net Job Creation`, na.rm = TRUE),
            firms = mean(`Number of Firms`, na.rm = TRUE),
            establishments = mean(`Number of Firms`, na.rm = TRUE),
            births = mean(`Establishment Births`, na.rm = TRUE),
            deaths = mean(`Establishment, Firm Deaths`, na.rm = TRUE),
            patents = mean(`Number of Utility Patents`, na.rm = TRUE),
            se = mean(`Proportion of S&E`, na.rm = TRUE),
            investments = mean(`Number of Firms Receiving VC`, na.rm = TRUE))

summary_variables <- c("employment", "sector_employment", "jobs", "firms",
                      "establishments", "births", "deaths", "patents", "se", "investments")
```

```

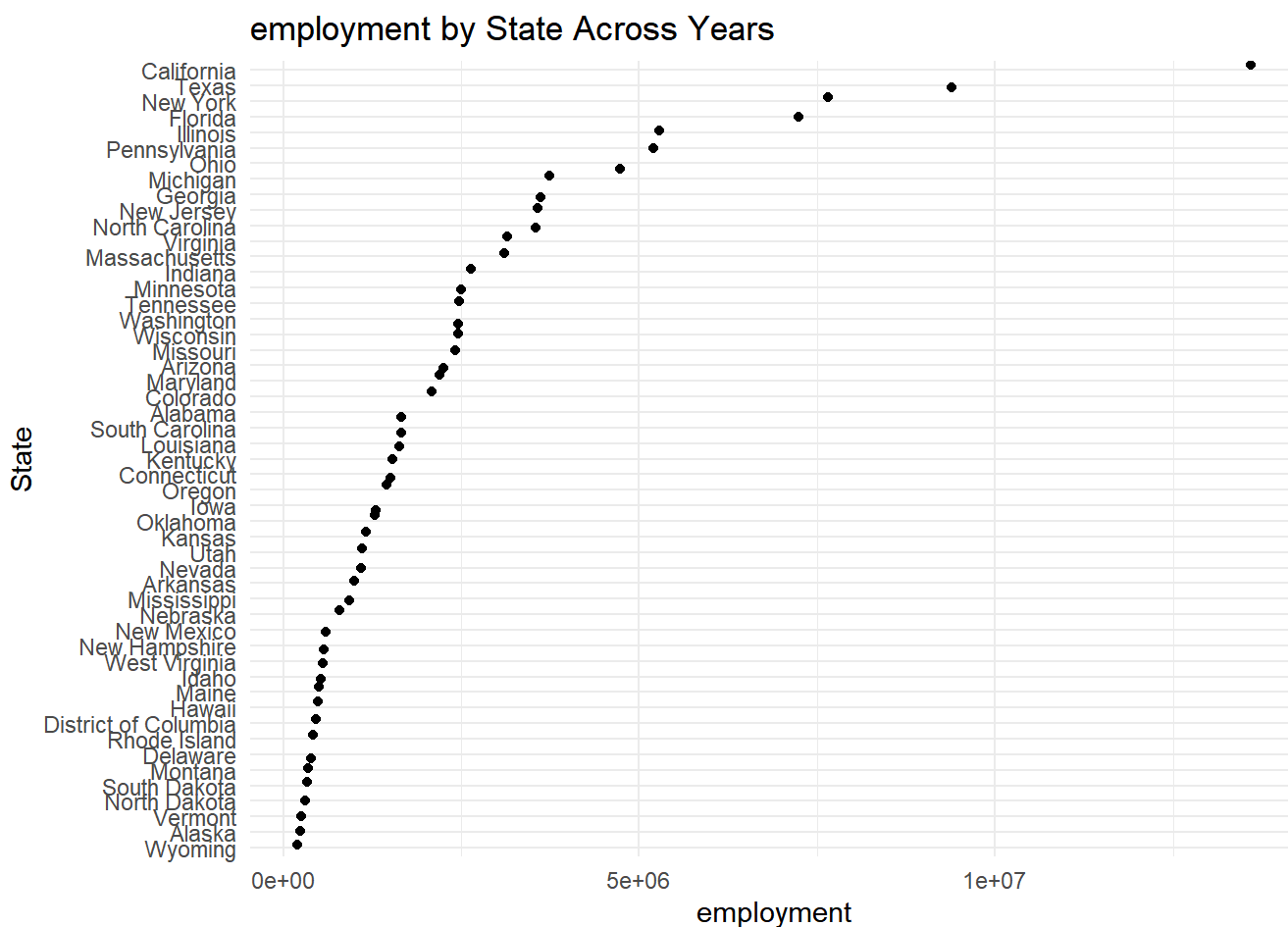
for (v in summary_variables) {
  summary_filtered <- full_summary |>
    select(State, all_of(v)) # Use all_of to ensure correct variable selection

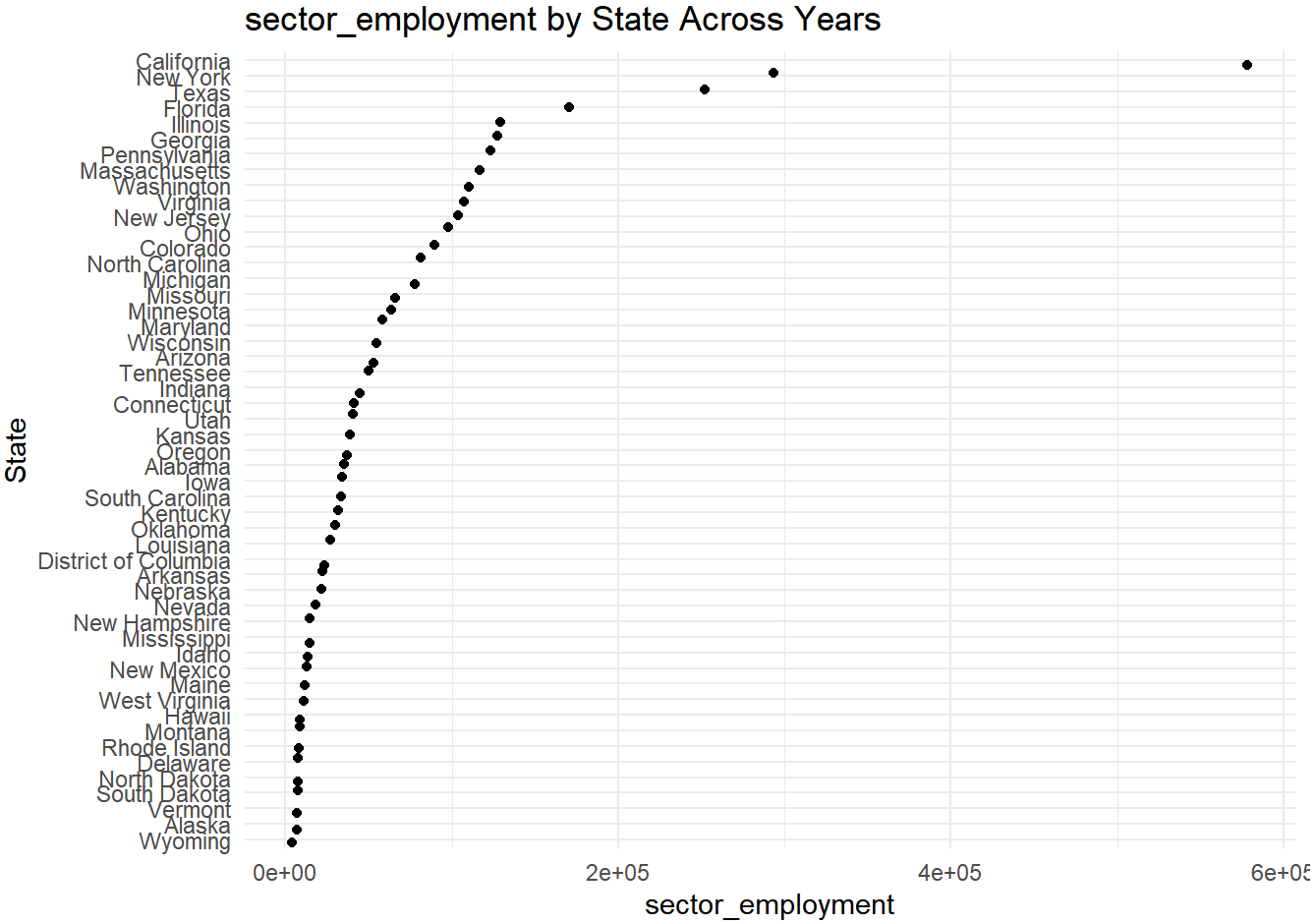
  # Reorder State by the values of the variable v from least to most
  summary_filtered$State <- fct_reorder(summary_filtered$State, summary_filtered[[v]])

  plot <- ggplot(summary_filtered, aes(x = State, y = .data[[v]])) +
    geom_jitter() +
    labs(title = paste(v, "by State Across Years"),
         x = "State",
         y = v) +
    theme_minimal() +
    coord_flip() # Optional for readability

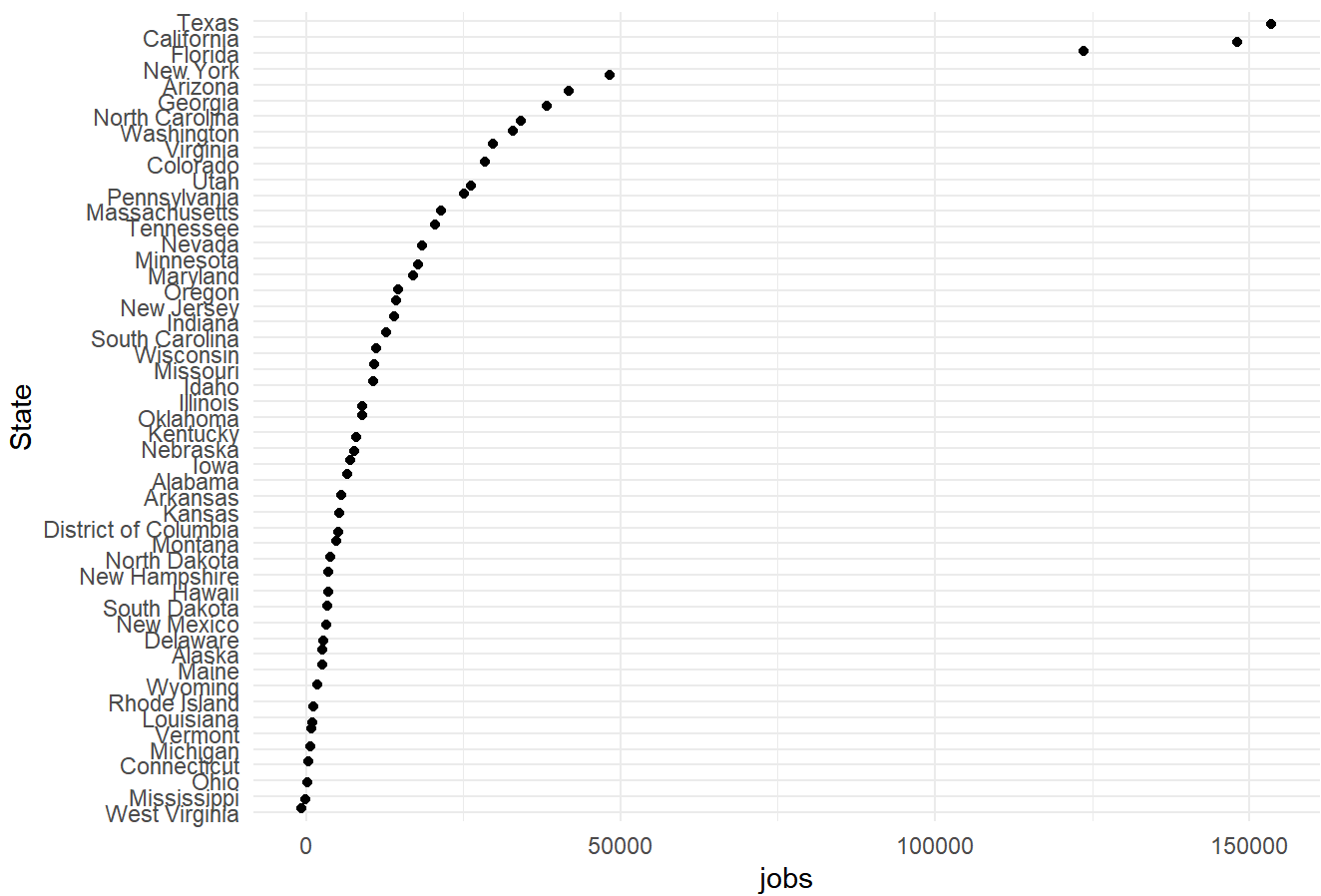
  print(plot)
}

```

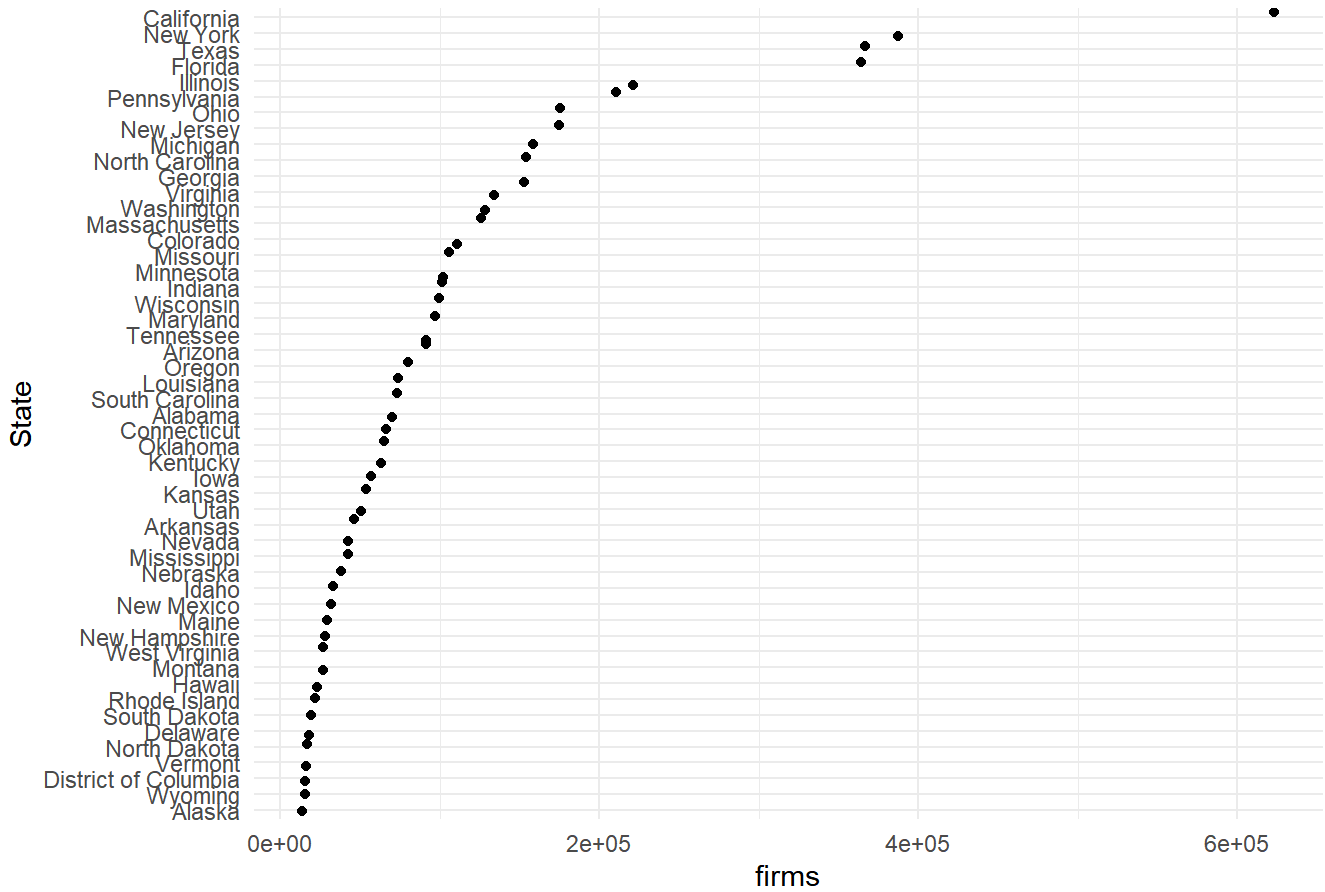




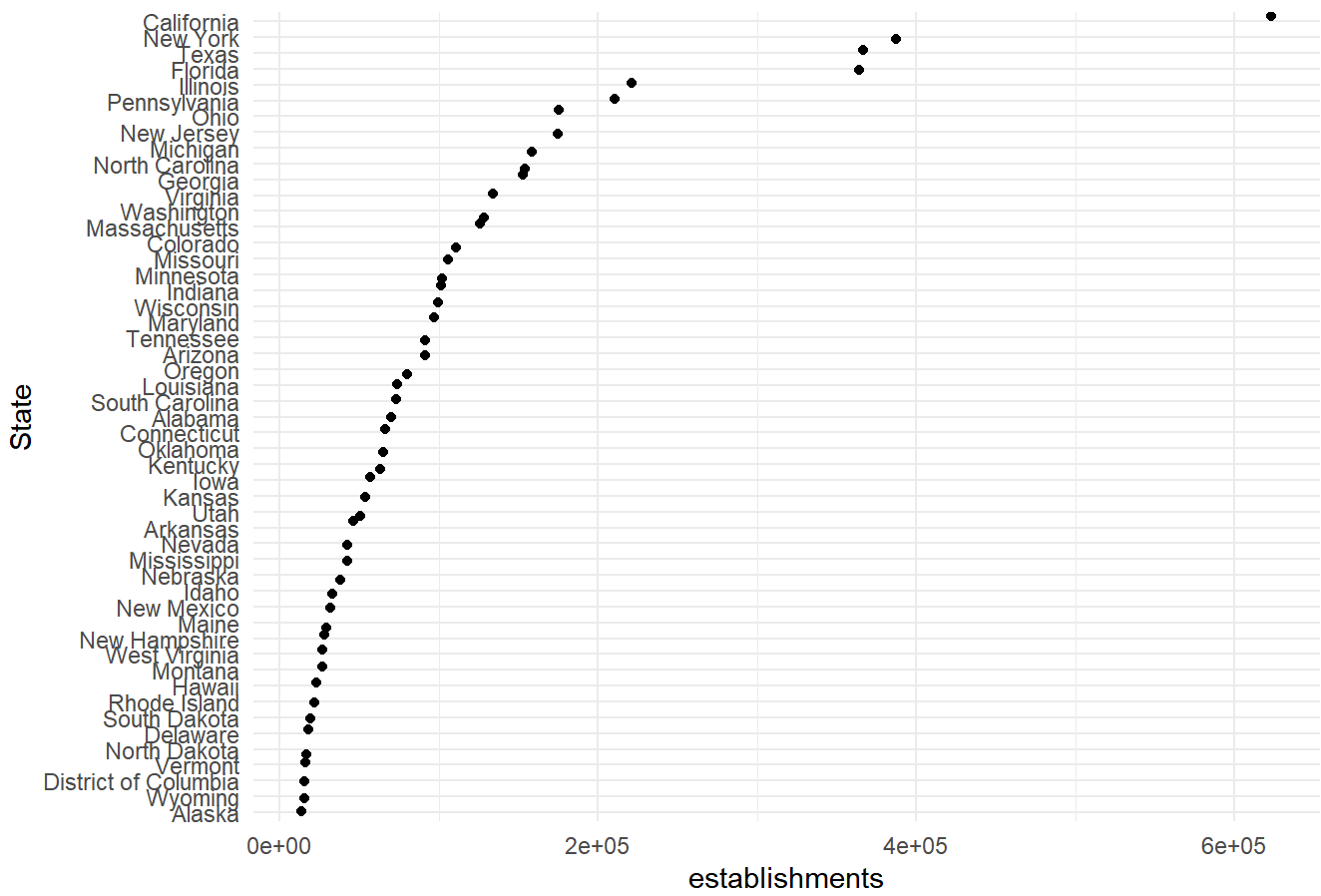
jobs by State Across Years



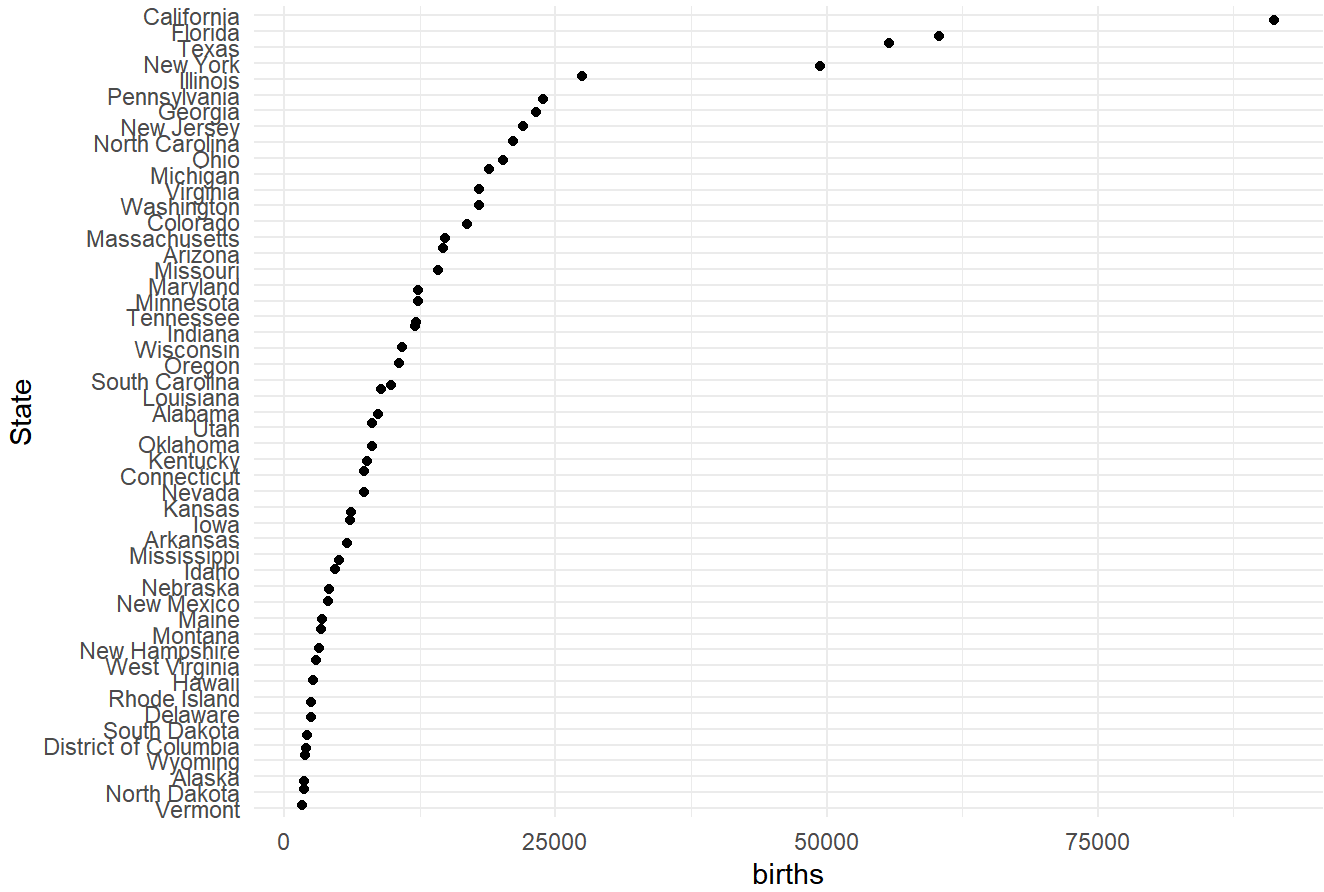
firms by State Across Years



establishments by State Across Years

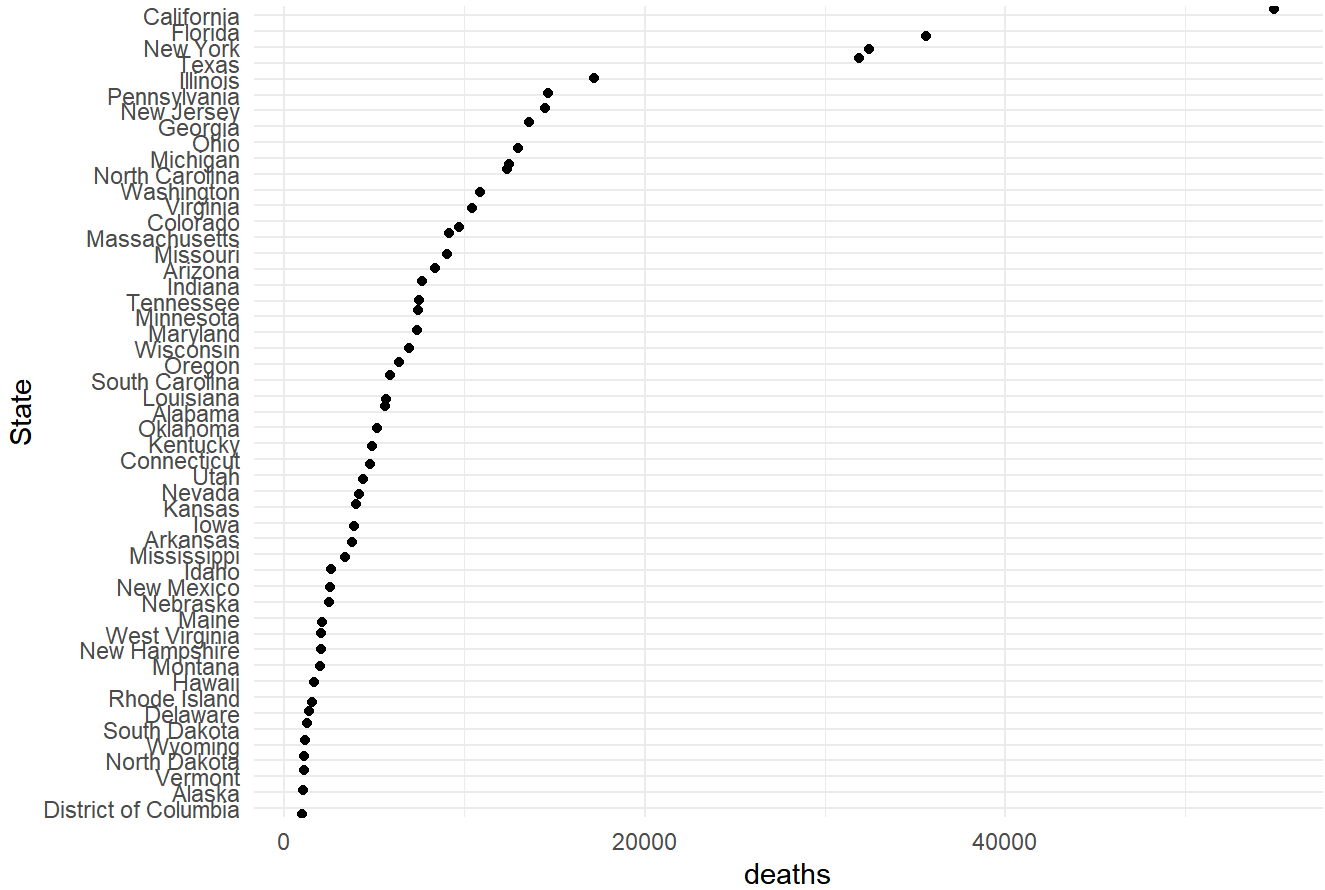


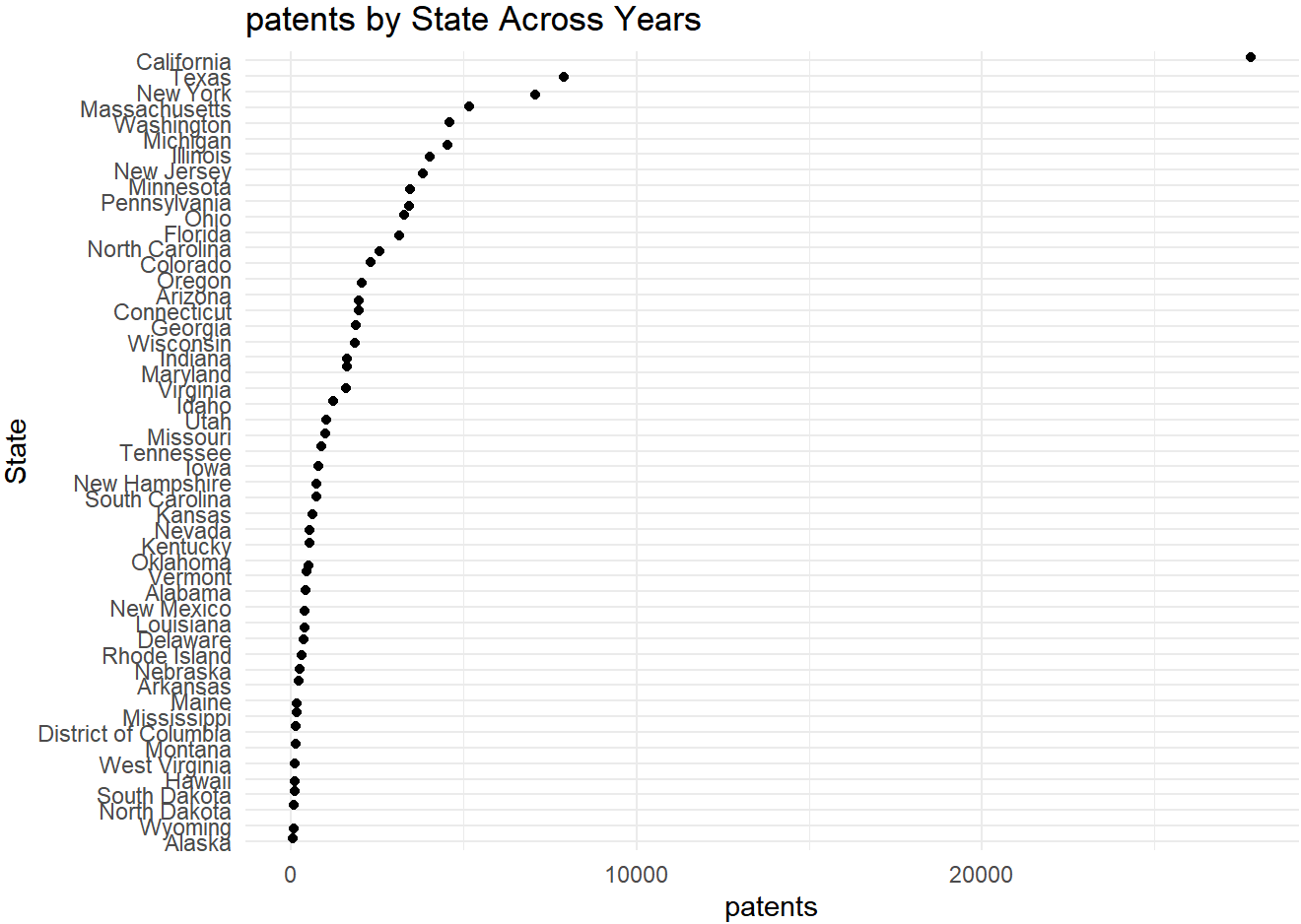
births by State Across Years



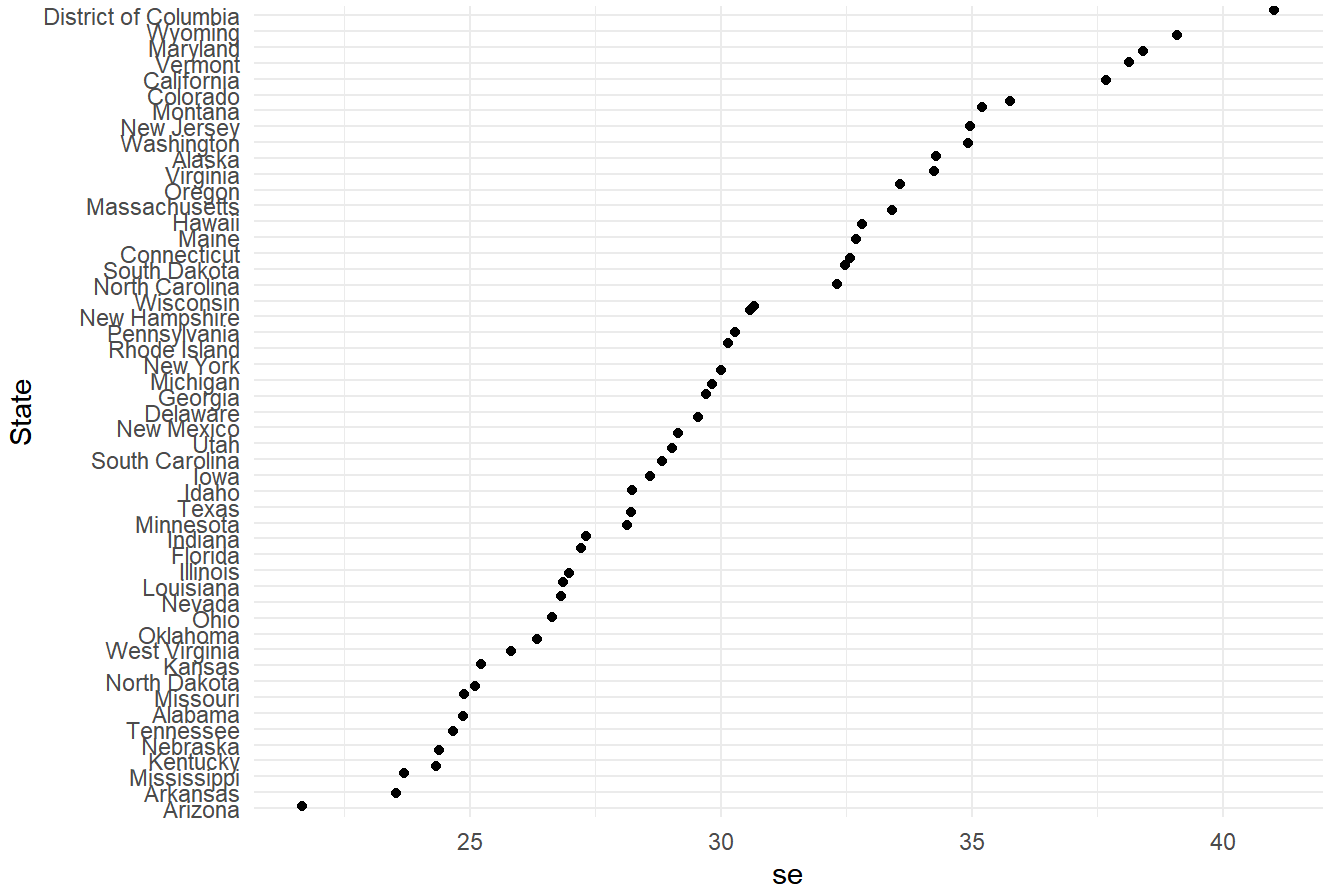


deaths by State Across Years





se by State Across Years



investments by State Across Years

