STAT350: Final Project

David Huu Pham / dhpham@sfu.ca / 301318482

```
death.csv.raw <- read_csv("data/proj/death.csv")
incd.csv.raw <- read_csv("data/proj/incd.csv")

head(death.csv.raw)
```

```
## # A tibble: 6 x 11
##   County  FIPS ‘Met Objective ~ ‘Age-Adjusted D~ ‘Lower 95% Conf~
##   <chr>  <dbl> <chr>                       <dbl>            <dbl>
## 1 Unite~     0 No                             46             45.9
## 2 Perry~ 21193 No                            126.            109.
## 3 Powel~ 21197 No                            125.            100.
## 4 North~  2185 No                            125.             73
## 5 Owsle~ 21189 No                            118.             83.1
## 6 Union~ 12125 No                            114.             89.9
## # ... with 6 more variables: ‘Upper 95% Confidence Interval for Death
## #   Rate‘ <dbl>, ‘Average Deaths per Year‘ <dbl>, ‘Recent Trend (2)‘ <chr>,
## #   ‘Recent 5-Year Trend (2) in Death Rates‘ <chr>, ‘Lower 95% Confidence
## #   Interval for Trend‘ <chr>, ‘Upper 95% Confidence Interval for Trend‘ <chr>
```

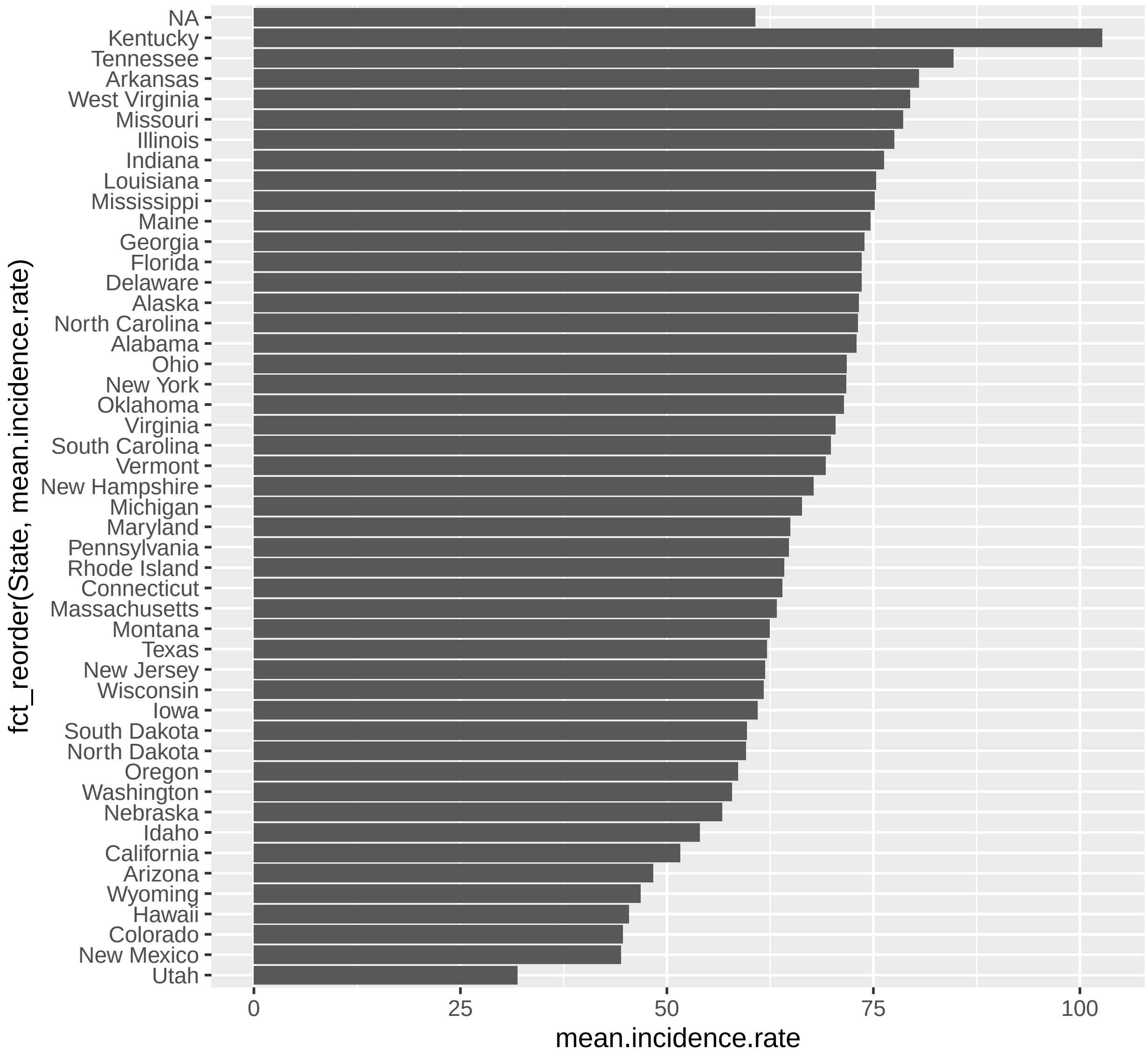
```
head(incd.csv.raw)
```

```
## # A tibble: 6 x 10
##   County  FIPS ‘Age-Adjusted I~ ‘Lower 95% Conf~ ‘Upper 95% Conf~
##   <chr>  <dbl> <chr>            <chr>            <chr>
## 1 US (S~     0 62.4             62.3             62.6
## 2 Autau~  1001 74.9             65.1             85.7
## 3 Baldw~  1003 66.9             62.4             71.7
## 4 Barbo~  1005 74.6             61.8             89.4
## 5 Bibb ~  1007 86.4             71               104.2
## 6 Bloun~  1009 69.7             61.2             79
## # ... with 5 more variables: ‘Average Annual Count‘ <chr>, ‘Recent
## #   Trend‘ <chr>, ‘Recent 5-Year Trend in Incidence Rates‘ <chr>, ‘Lower 95%
## #   Confidence Interval_1‘ <chr>, ‘Upper 95% Confidence Interval_1‘ <chr>
```

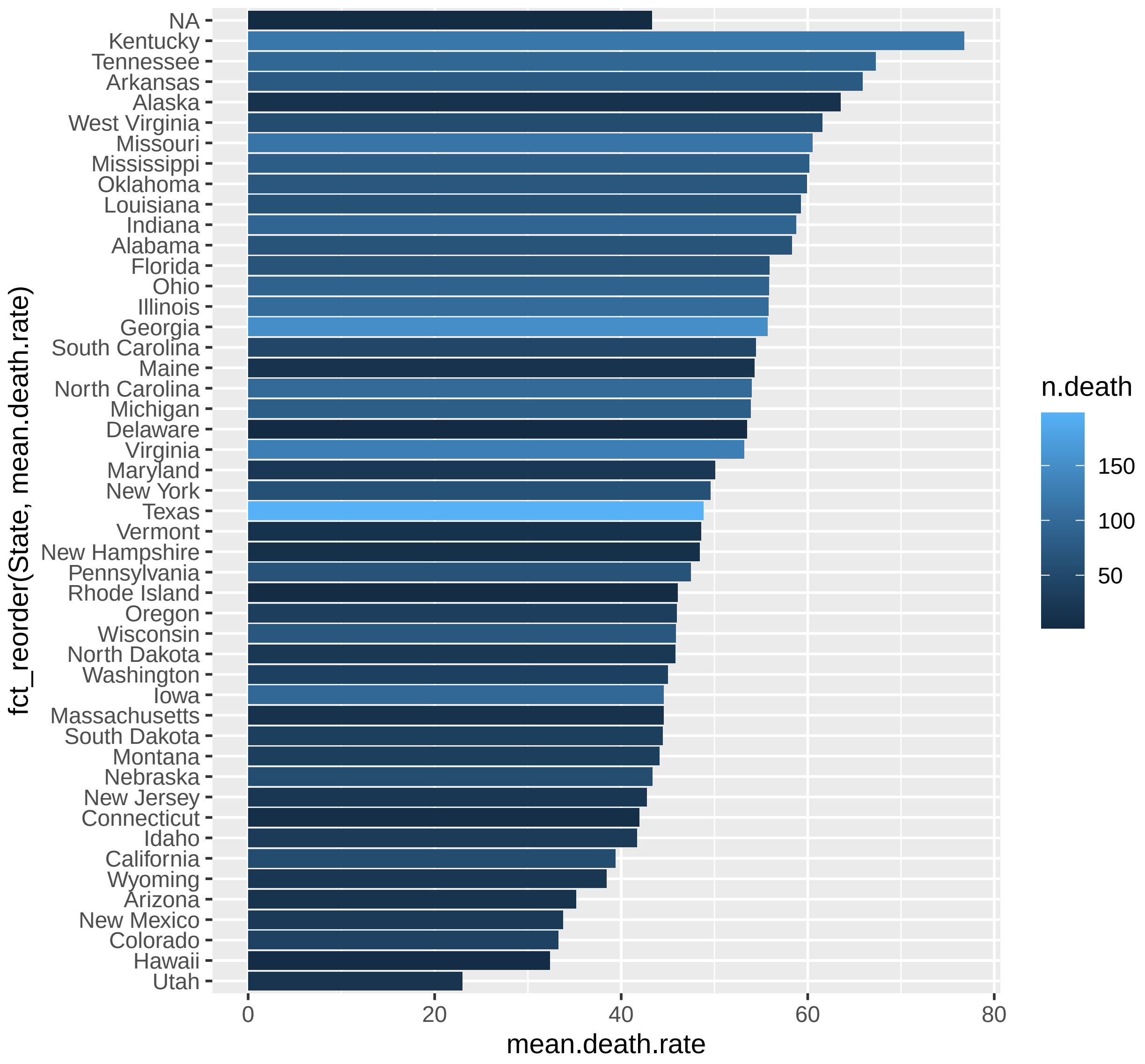
```
tidy_and_asnumeric <- function(.data, .cols.asnumeric) {
  .data %>%
    # Removes footnote numbers from column names and County names
    rename_with(~ str_remove(.x, "\\?? ?\\(.*\\)")) %>%
    mutate(County = str_remove_all(County, " ?\\(.*\\)")) %>%
    mutate(County = str_remove_all(County, " ?<.*>")) %>%
    # Creates a column for State, removes State from County
    separate(col=County, into=c("County", "State"), sep=", ") %>%
```

```r
    # Due to suppression,
      # incidence for counties with 3 or less counts not included
      # deaths for counties with 10 or less counts not included
    # As well, if less than 10? counts, the average trend is not computed
    # We take out all these suppressed cases, so we always have the average trend
    # filter(across(
    #   .cols = starts_with(.cols.rate),
    #   .fns = ~ !str_detect(.x, "\\*") & !is.na(.x)
    # )) %>%
    # filter(`Recent Trend` != "*" | `Recent Trend` != "**") %>%

    # Convert State to factors
    mutate(State = as_factor(State)) %>%

    # Remove any non-numeric characters from the columns we want to convert to numeric
    mutate(across(
      .cols = starts_with(.cols.asnumeric),
      .fns = ~ str_remove_all(.x, "[^\\+\\-\\.[:digit:]]+")
    )) %>%
    # Convert columns to numeric
    mutate(across(
      .cols = starts_with(.cols.asnumeric),
      .fns = as.numeric
    )) %>%
    # Filter out any rows with missing rates data
    filter_at(
      vars(starts_with(.cols.asnumeric)),
      any_vars(!is.na(.))
    )
}

# Convert these columns to numeric data types
incd.cols.asnumeric <- c("Age-Adjusted", "Upper", "Lower", "Average", "Recent 5")
# Split data up based on rate and trends
incd.untidy.rates <- incd.csv.raw %>% select(1:5)
incd.untidy.trends <- incd.csv.raw %>% select(1:2, !(3:5))

incd.tidy.rates <- tidy_and_asnumeric(incd.untidy.rates, incd.cols.asnumeric)
incd.tidy.trends <- tidy_and_asnumeric(incd.untidy.trends, incd.cols.asnumeric) %>%
  rename("Average Incidence Counts per Year" = "Average Annual Count",
         "Lower 95% Confidence Interval for Trend in Incidence Rate" = "Lower 95% Confidence Interval_1
         "Upper 95% Confidence Interval for Trend in Incidence Rate" = "Upper 95% Confidence Interval_1


# Convert these columns to numeric data types
death.cols.asnumeric <- c("Upper", "Lower", "Recent 5")
# Convert these columns to logical data types
death.cols.aslogical <- c("Met Objective")
death.csv.aslogical <- death.csv.raw %>%
  mutate(across(
    .cols = starts_with(death.cols.aslogical),
    .fns = str_detect,
    pattern = fixed("yes", ignore_case=TRUE)
  ))

death.untidy.rates <- death.csv.aslogical %>% select(1:6)
death.untidy.trends <- death.csv.aslogical %>% select(1:3, 7:last_col())

death.tidy.rates <- tidy_and_asnumeric(death.untidy.rates, death.cols.asnumeric)
death.tidy.trends <- tidy_and_asnumeric(death.untidy.trends, death.cols.asnumeric) %>%
  rename_with(~ str_replace(.x, "Interval for Trend$", "Interval for Trend in Death Rates"))
```

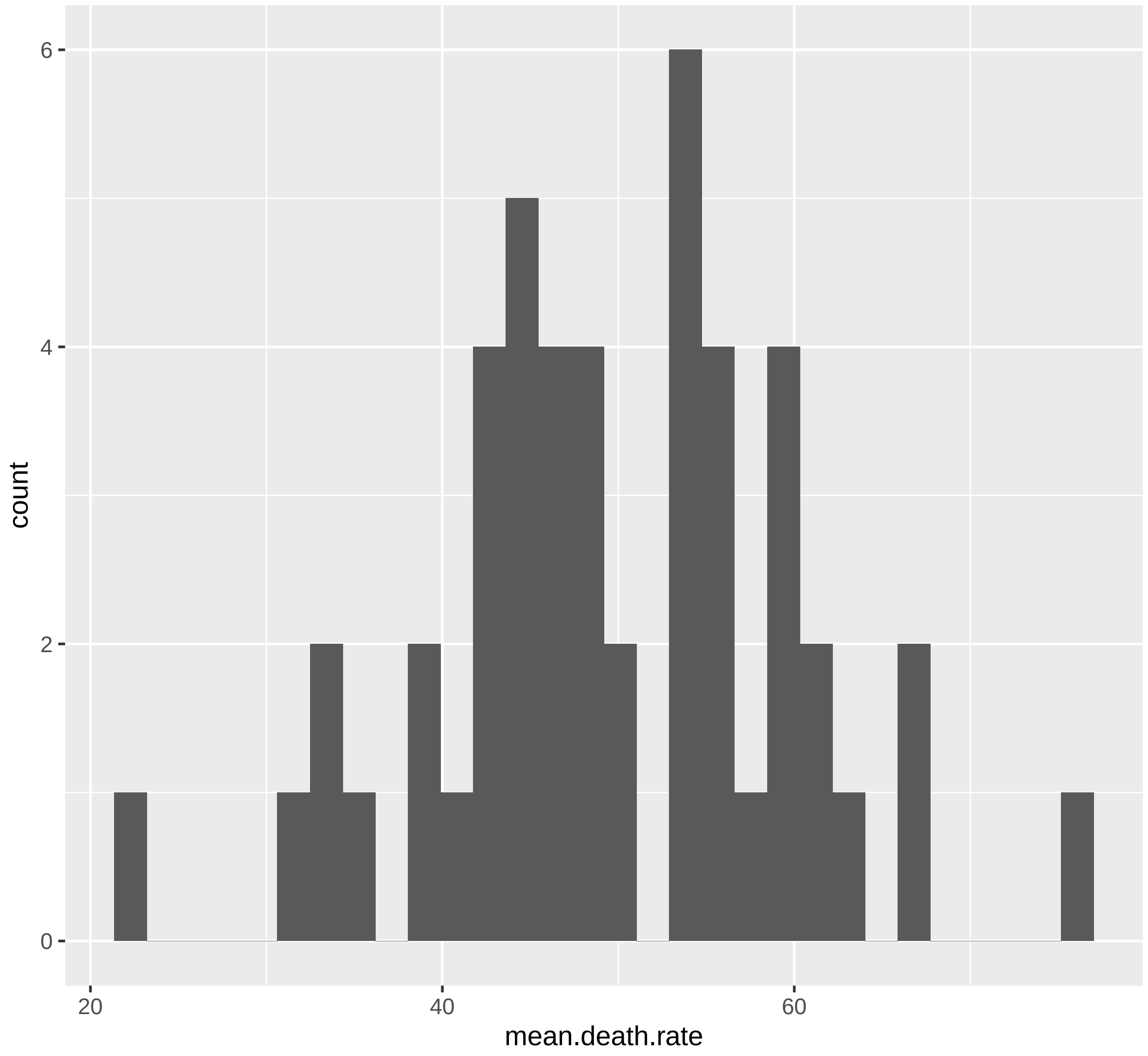
```
incd.rates.bystate <- incd.tidy.rates %>%
  group_by(State) %>%
  summarize(
    n.incd = n(),
    mean.incidence.rate = mean(‘Age-Adjusted Incidence Rate - cases per 100,000‘)
  )

death.rates.bystate <- death.tidy.rates %>%
  group_by(State) %>%
  summarize(
    n.death = n(),
    mean.death.rate = mean(‘Age-Adjusted Death Rate‘)
  )

rates.bystate <- incd.rates.bystate %>%
  inner_join(death.rates.bystate, by="State")

ggplot(rates.bystate, aes(x=mean.incidence.rate, y=fct_reorder(State, mean.incidence.rate))) +
  geom_col()
```



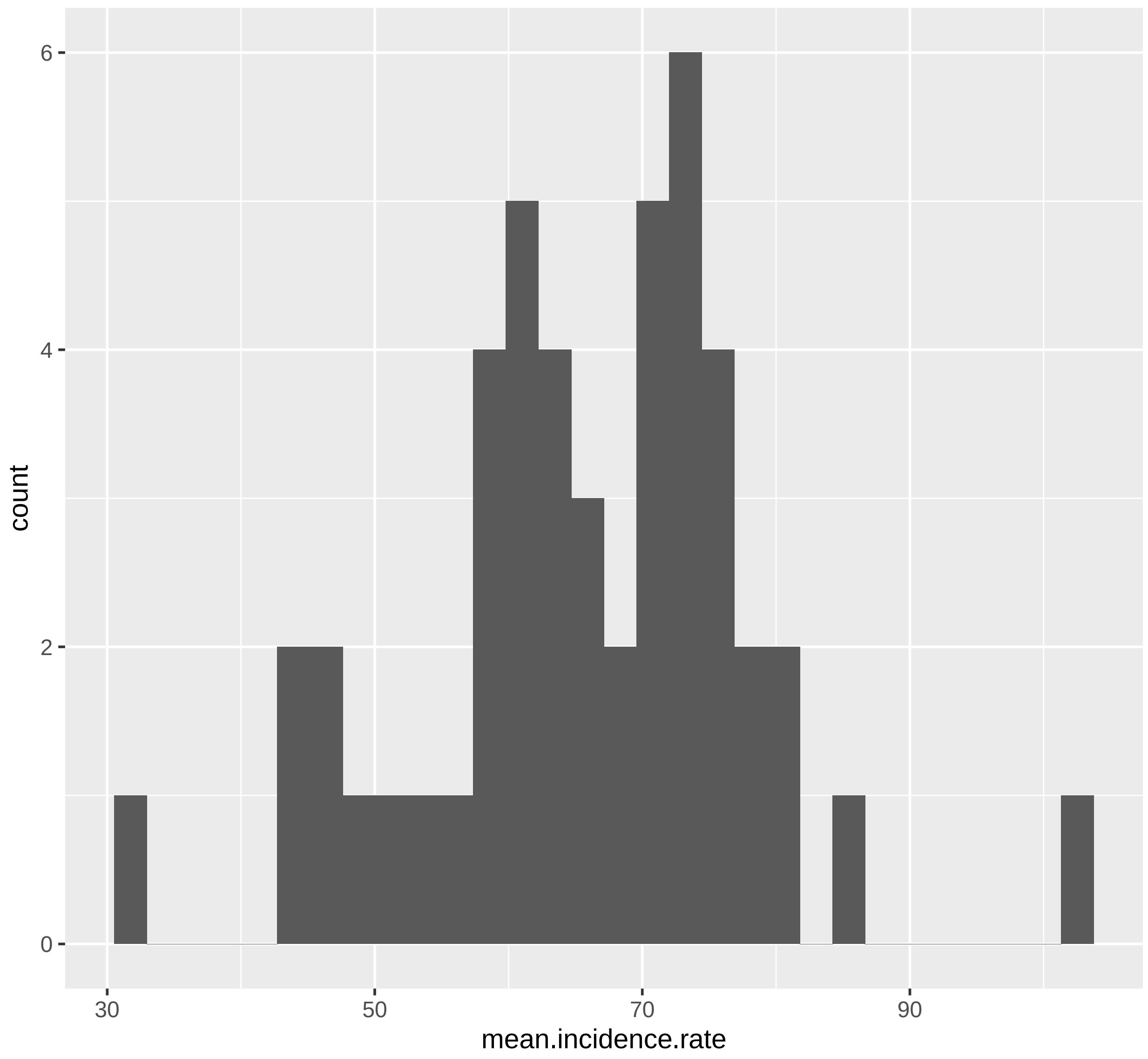
```
ggplot(rates.bystate, aes(x=mean.death.rate, y=fct_reorder(State, mean.death.rate))) +
  geom_col(aes(fill = n.death))
```



```
ggplot(rates.bystate, aes(x=mean.death.rate)) +
  geom_histogram()
```



```
ggplot(rates.bystate, aes(x=mean.incidence.rate)) +
  geom_histogram()
```


count
6
4
2
0
30
50
70
90
mean.incidence.rate

“‘