

# truancy

Jean-Luc Jackson

3/30/2022

```
# Load libraries
library(tidyverse)
```

```
# Load data
data_list <- readRDS("../datasets/awesome_data.rds")
names(data_list)
```

```
## [1] "attendance"          "student_teacher_ratios" "teacher_salaries"
## [4] "free_and_reduced_lunch" "school_code"          "consolidated"
```

## Load Data

```
attendance <- data_list$attendance
stu_teach <- data_list$student_teacher_ratios
teach_demo <- data_list$teacher_demographics
teach_sals <- data_list$teacher_salaries
lunches <- data_list$free_and_reduced_lunch

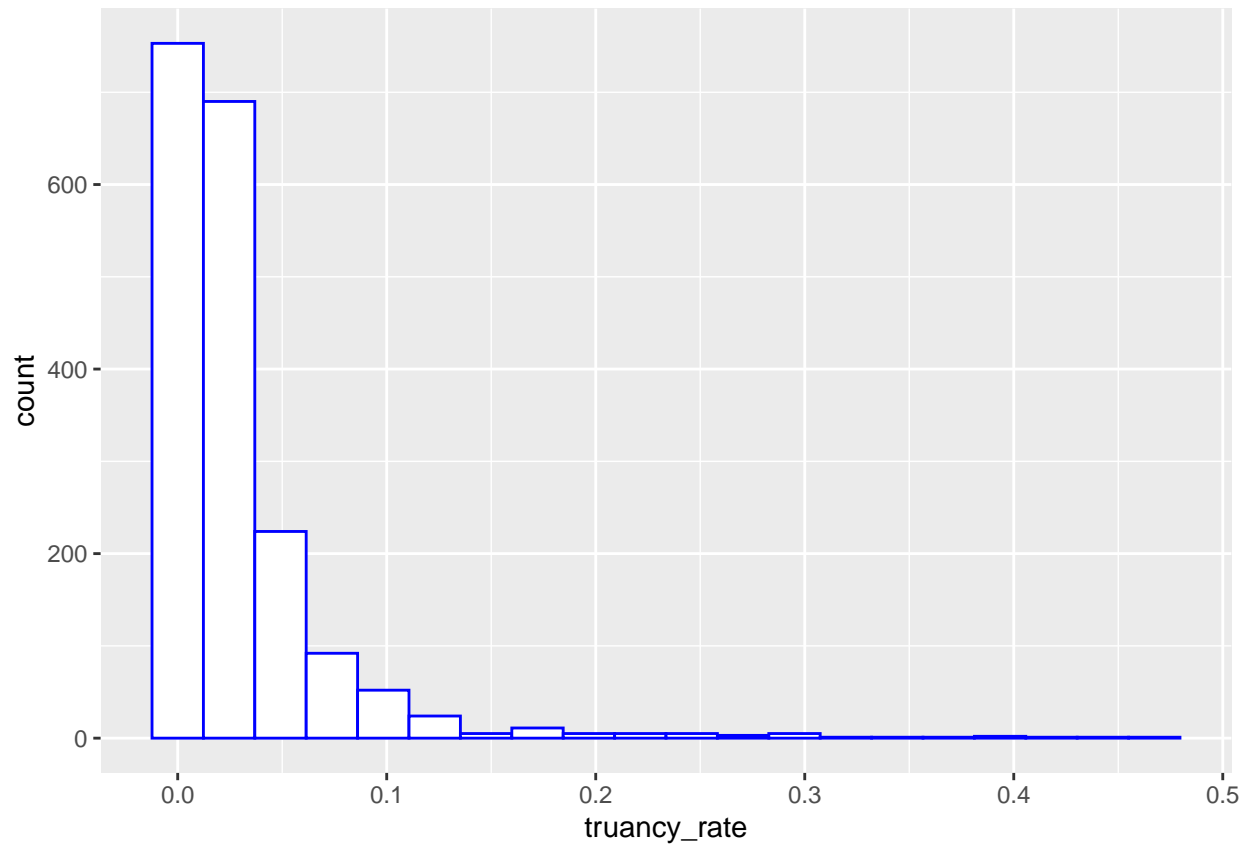
consol <- data_list$consolidated
```

## Attendance

```
# Look at attendance
names(attendance)
```

```
## [1] "school_year"          "county_code"
## [3] "county_name"          "district_code"
## [5] "district_name"        "school_code"
## [7] "school_name"          "enrollment"
## [9] "length_of_school_year" "days_possible_attendance"
## [11] "days_attended"        "days_excused_absence"
## [13] "days_unexcused_absent" "attendance_rate"
## [15] "truancy_rate"
```

```
attendance %>%
  ggplot() +
  aes(x = truancy_rate) +
  geom_histogram(color = 'blue',
                 fill = 'white',
                 bins = 20)
```



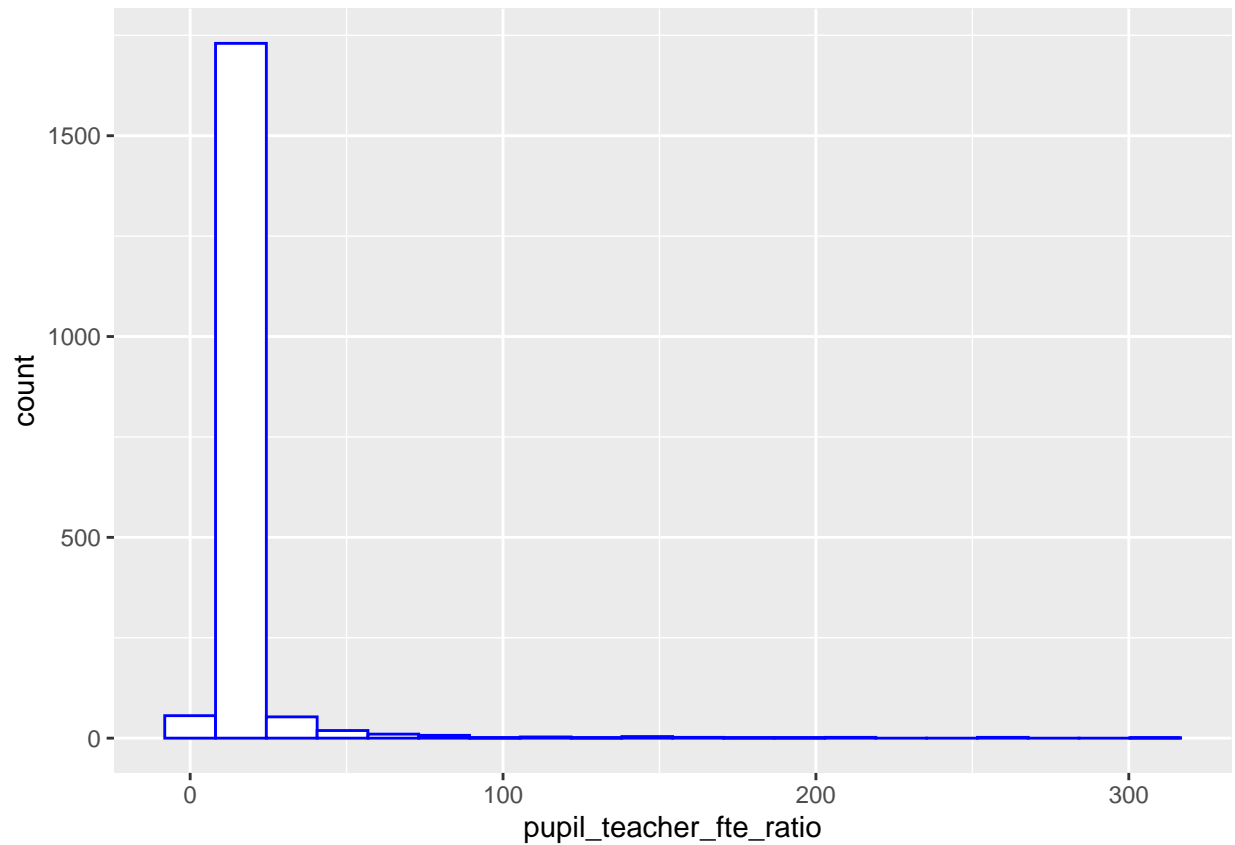
## Student-Teacher Ratios

```
# Student Teacher Ratios
```

```
names(stu_teach)
```

```
## [1] "county_code"      "county_name"
## [3] "district_code"    "district_name"
## [5] "school_code"      "school_name"
## [7] "pk_12_count"      "teacher_fte"
## [9] "pupil_teacher_fte_ratio"
```

```
stu_teach %>%
  ggplot() +
  aes(x = pupil_teacher_fte_ratio) +
  geom_histogram(bins = 20,
                 color = 'blue',
                 fill = 'white')
```



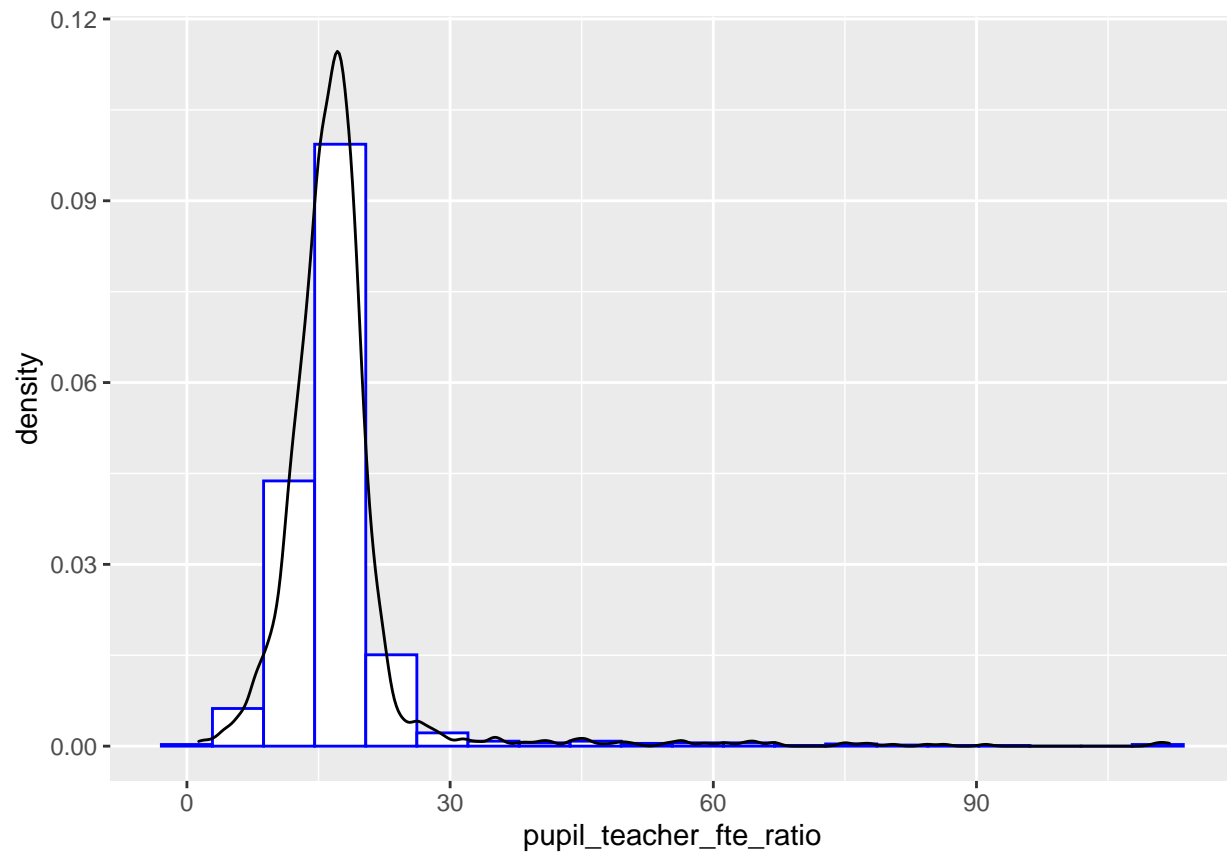
Defining FTE: <https://www.edglossary.org/student-teacher-ratio/>

Preschools are outliers – ratios up to \*300 to 1\*\*

```
stu_teach %>%
  filter(pupil_teacher_fte_ratio < 120) %>%
  summarize(n = n())
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1879
```

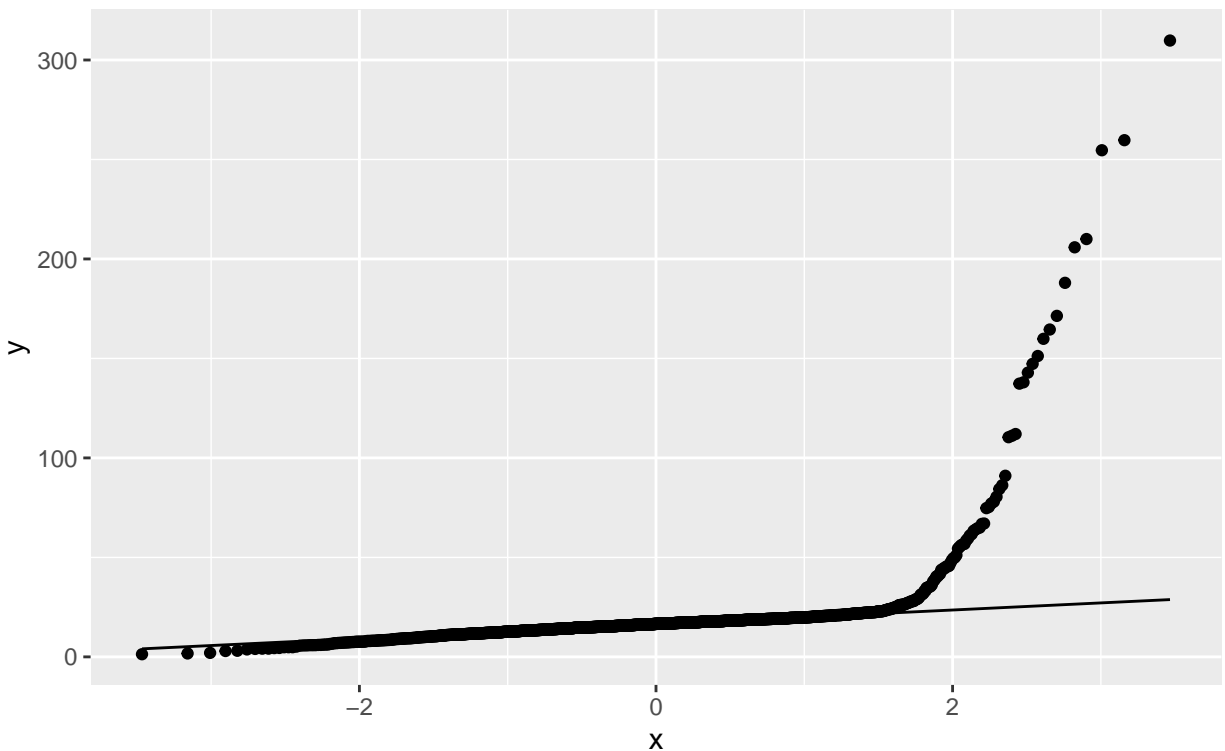
```
stu_teach %>%
  filter(pupil_teacher_fte_ratio < 120) %>%
  ggplot() +
  aes(x = pupil_teacher_fte_ratio, y = ..density..) +
  geom_histogram(bins = 20,
                 color = 'blue',
                 fill = 'white') +
  geom_density()
```



```
stu_teach %>%  
  ggplot() +  
    aes(sample = pupil_teacher_fte_ratio) +  
    stat_qq() + stat_qq_line() +  
    labs(  
      title = "Q-Q Plot",  
      subtitle = "Checking normality"  
    )
```

## Q-Q Plot

Checking normality



very heavy tail. maybe we don't include preschools? and just focus on grade schools

## Teacher Salaries

```
# Teacher Salaries
names(teach_sals)

## [1] "organization_code"      "organization_name"      "charter_fte"
## [4] "charter_average_salary" "non_charter_fte"        "non_charter_salary"
## [7] "all_fte"                "all_average"

# Make numeric columns
teach_salaries <- teach_sals %>%
  transform(charter_average_salary = as.numeric(charter_average_salary),
            non_charter_salary = as.numeric(non_charter_salary),
            all_average = as.numeric(all_average)) %>%
  mutate(
    charter_flag = charter_average_salary > 0
  )

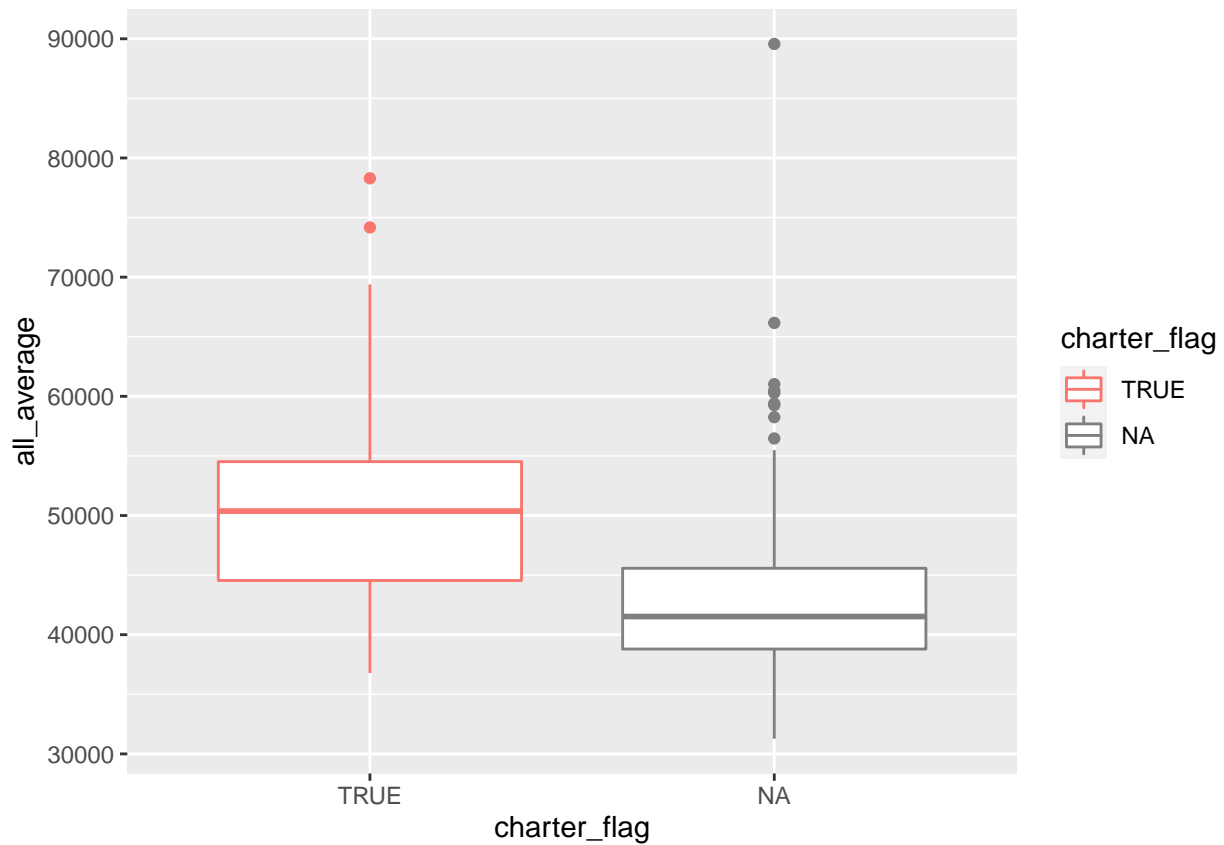
teach_salaries %>%
  distinct() %>%
  summarize(n = n())

##      n
## 1 199
```

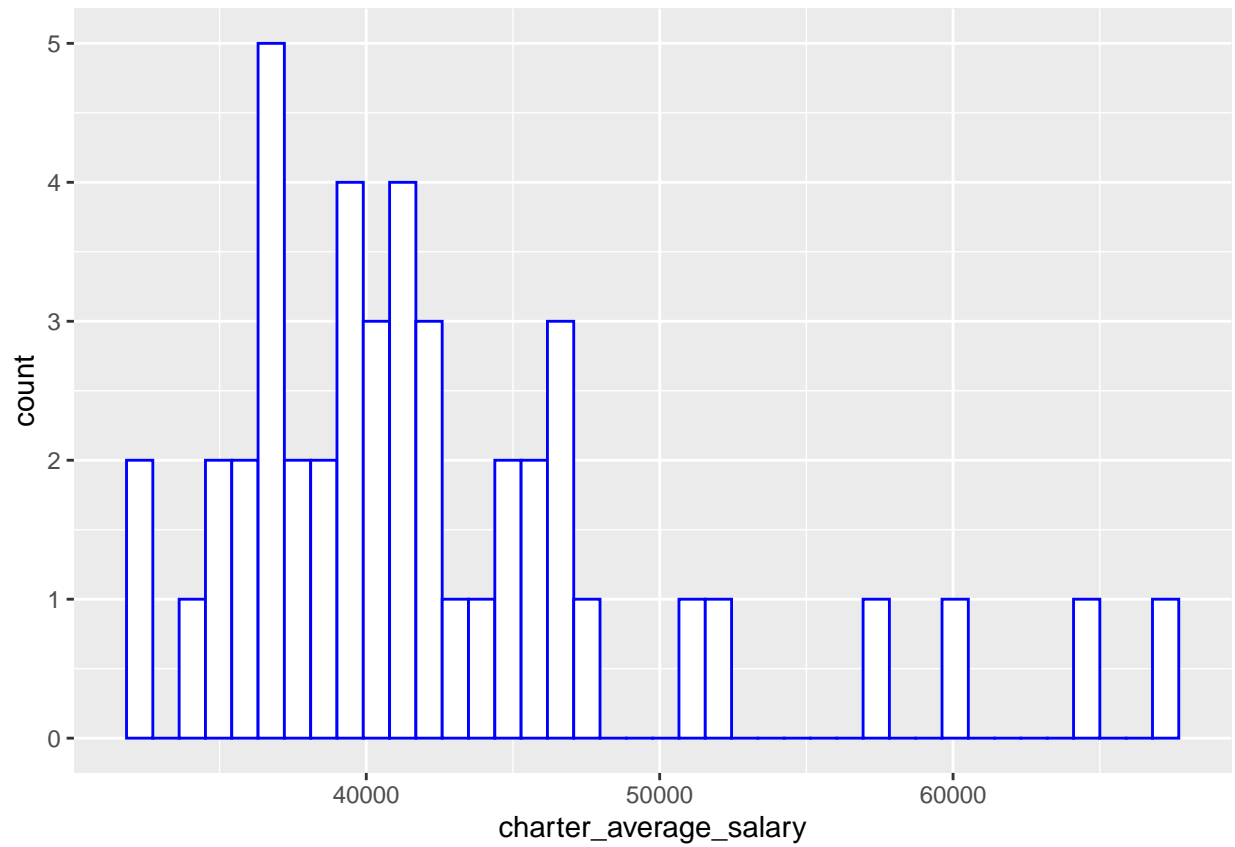
```
teach_salaries %>%
  group_by(charter_flag) %>%
  summarize(n = n())
```

```
## # A tibble: 2 x 2
##   charter_flag     n
##   <lgl>         <int>
## 1 TRUE           46
## 2 NA            153
```

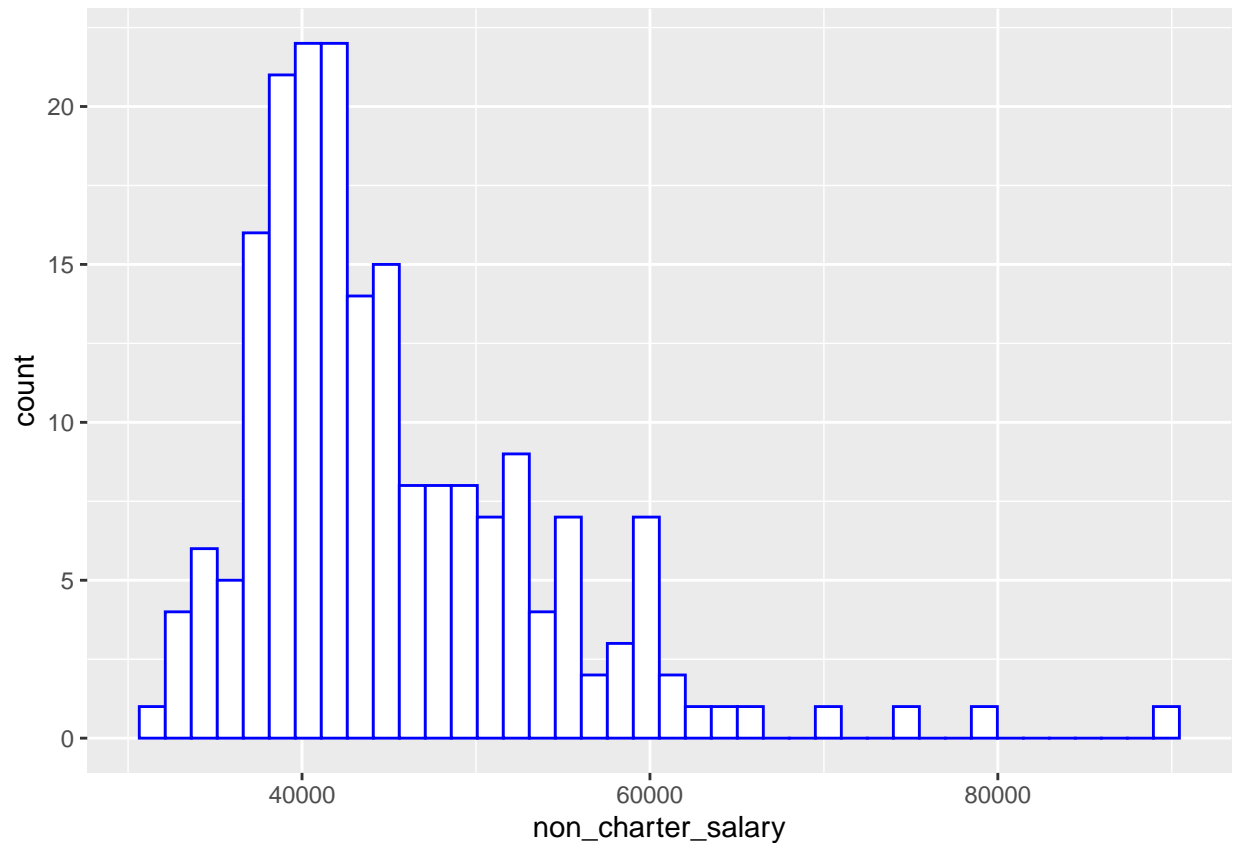
```
teach_salaries %>%
  ggplot() +
  aes(x = charter_flag,
       y = all_average,
       color = charter_flag) +
  geom_boxplot()
```



```
teach_salaries %>%
  filter(charter_average_salary > 0) %>%
  ggplot() +
  aes(x = charter_average_salary) +
  geom_histogram(bins = 40,
                 color = 'blue',
                 fill = 'white')
```



```
teach_salaries %>%  
  filter(non_charter_salary > 0) %>%  
  ggplot() +  
  aes(x = non_charter_salary) +  
  geom_histogram(bins = 40,  
                 color = 'blue',  
                 fill = 'white')
```



```
# t-test of charter vs non-charter salaries
charter_salaries <- teach_salaries %>%
  filter(charter_average_salary > 0) %>%
  select(charter_average_salary)

non_charter_salaries <- teach_salaries %>%
  filter(non_charter_salary > 0) %>%
  select(non_charter_salary)

t.test(charter_salaries$charter_average_salary,
        non_charter_salaries$non_charter_salary)

##
## Welch Two Sample t-test
##
## data: charter_salaries$charter_average_salary and non_charter_salaries$non_charter_salary
## t = -2.1164, df = 74.114, p-value = 0.03767
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -5323.653 -160.558
## sample estimates:
## mean of x mean of y
## 42535.95 45278.05

mean(charter_salaries$charter_average_salary)

## [1] 42535.95
```



```

mean(non_charter_salaries$non_charter_salary)

## [1] 45278.05

# Joining into consolidated df
consolidated_salaries <- teach_salaries %>%
  select(organization_code,
         charter_average_salary,
         non_charter_salary,
         all_average) %>%
  left_join(consol, .,
            by = c("district_code" = "organization_code"))

consolidated_salaries %>%
  select(everything()) %>% # replace to your needs
  summarise_all(funs(sum(is.na(.))))

## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.

## # A tibble: 1 x 17
##   school_year county_code county_name district_code district_name school_code
##   <int>      <int>      <int>      <int>      <int>      <int>
## 1         0         0         0         0         0         0
## # ... with 11 more variables: school_name <int>, days_unexcused_absent <int>,
## #   truancy_rate <int>, pupil_teacher_fte_ratio <int>,
## #   percent_free_and_reduced <int>, lunch_subsidy_eligibility <int>,
## #   average_salary <int>, charter <int>, charter_average_salary <int>,
## #   non_charter_salary <int>, all_average <int>

```

## Free plus Reduced Lunches

```

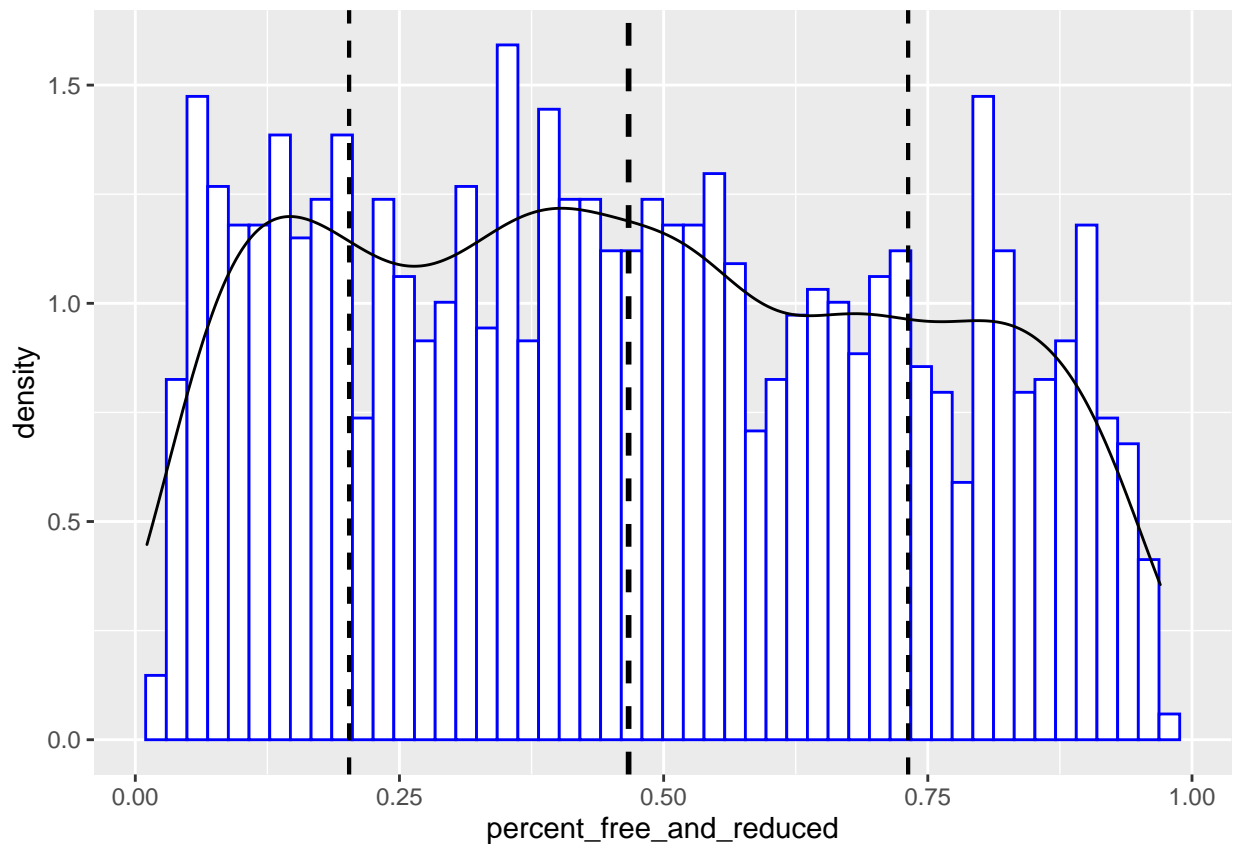
# Free & Reduced Lunch
names(lunches)

## [1] "county_code"      "county_name"
## [3] "district_code"    "district_name"
## [5] "school_code"      "school_name"
## [7] "pk_12_count"      "free_lunch"
## [9] "reduced_lunch"     "not_eligible"
## [11] "free_and_reduced_count" "percent_free"
## [13] "percent_reduced"   "percent_free_and_reduced"
## [15] "lunch_subsidy_eligibility"

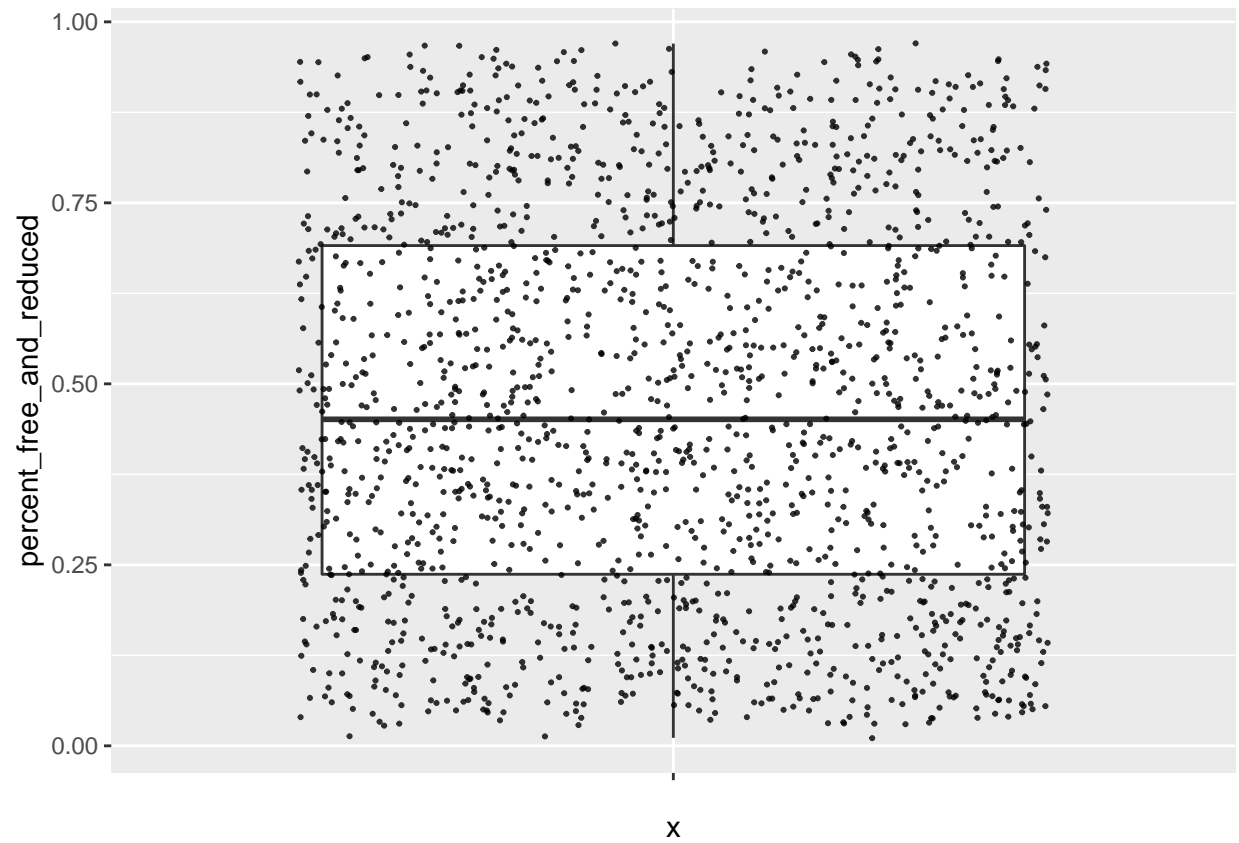
```

```
lunches %>%
  ggplot() +
    aes(x = percent_free_and_reduced,
        y = ..density..) +
    geom_histogram(bins = 50,
                  color = 'blue',
                  fill = 'white') +
    geom_vline(aes(xintercept = mean(lunches$percent_free_and_reduced)), size = 1.0, linetype = "dashed")
    geom_vline(aes(xintercept = (mean(lunches$percent_free_and_reduced) + sd(lunches$percent_free_and_reduced))), size = 1.0, linetype = "dashed")
    geom_vline(aes(xintercept = (mean(lunches$percent_free_and_reduced) - sd(lunches$percent_free_and_reduced))), size = 1.0, linetype = "dashed")
    geom_density()
```

```
## Warning: Use of `lunches$percent_free_and_reduced` is discouraged. Use `percent_free_and_reduced` instead.
## Use of `lunches$percent_free_and_reduced` is discouraged. Use `percent_free_and_reduced` instead.
## Use of `lunches$percent_free_and_reduced` is discouraged. Use `percent_free_and_reduced` instead.
## Use of `lunches$percent_free_and_reduced` is discouraged. Use `percent_free_and_reduced` instead.
## Use of `lunches$percent_free_and_reduced` is discouraged. Use `percent_free_and_reduced` instead.
```



```
lunches %>%
  ggplot() +
    aes(x = "", y = percent_free_and_reduced) +
    geom_boxplot() +
    geom_jitter(color = 'black', size = 0.4, alpha = 0.8)
```



```
quantile(lunches$percent_free_and_reduced)
```

```
##      0%   25%   50%   75%  100%  
## 0.011 0.237 0.451 0.691 0.970
```