YuHW00ST430

Haozhe (Jerry) Yu

2023-09-07

Question 1

> The dataset sat.txt comes from a study entitled "Getting What You Pay For: The Debate Over Equity in Public School Expenditures.'' Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Make sure you write something about anything you choose to present. You are not expected to make any substantive conclusions from the data. Do some short numerical summaries of the data, commenting on any features that you find interesting. (Please follow the R Codes given in Prostate Data and Pima Data)

First we input the SAT data from a local txt file. We did this because the sat.txt from the pdf had an error.

```
satnames <- read.delim2("Datasets/SAT.txt",
                        nrows = 1,
                        sep = "",
                        header = FALSE) %>%
  as.character()

ssatnames <- c("State", satnames)

sat <-  as_tibble (read.delim2(
  "Datasets/SAT.txt",
  header = TRUE,
  sep = "",
  col.names = ssatnames
))

sat$expend <- 1000 * as.numeric(sat$expend)
sat$salary <- 1000 * as.numeric(sat$salary)
sat$ratio <- as.numeric(sat$ratio)
sat
```

```
## # A tibble: 50 x 8
##    State       expend ratio salary takers verbal  math total
##    <chr>        <dbl> <dbl>  <dbl>  <int>  <int> <int> <int>
##  1 Alabama       4405  17.2  31144      8    491   538  1029
##  2 Alaska        8963  17.6  47951     47    445   489   934
##  3 Arizona       4778  19.3  32175     27    448   496   944
##  4 Arkansas      4459  17.1  28934      6    482   523  1005
##  5 California    4992  24    41078     45    417   485   902
##  6 Colorado      5443  18.4  34571     29    462   518   980
```

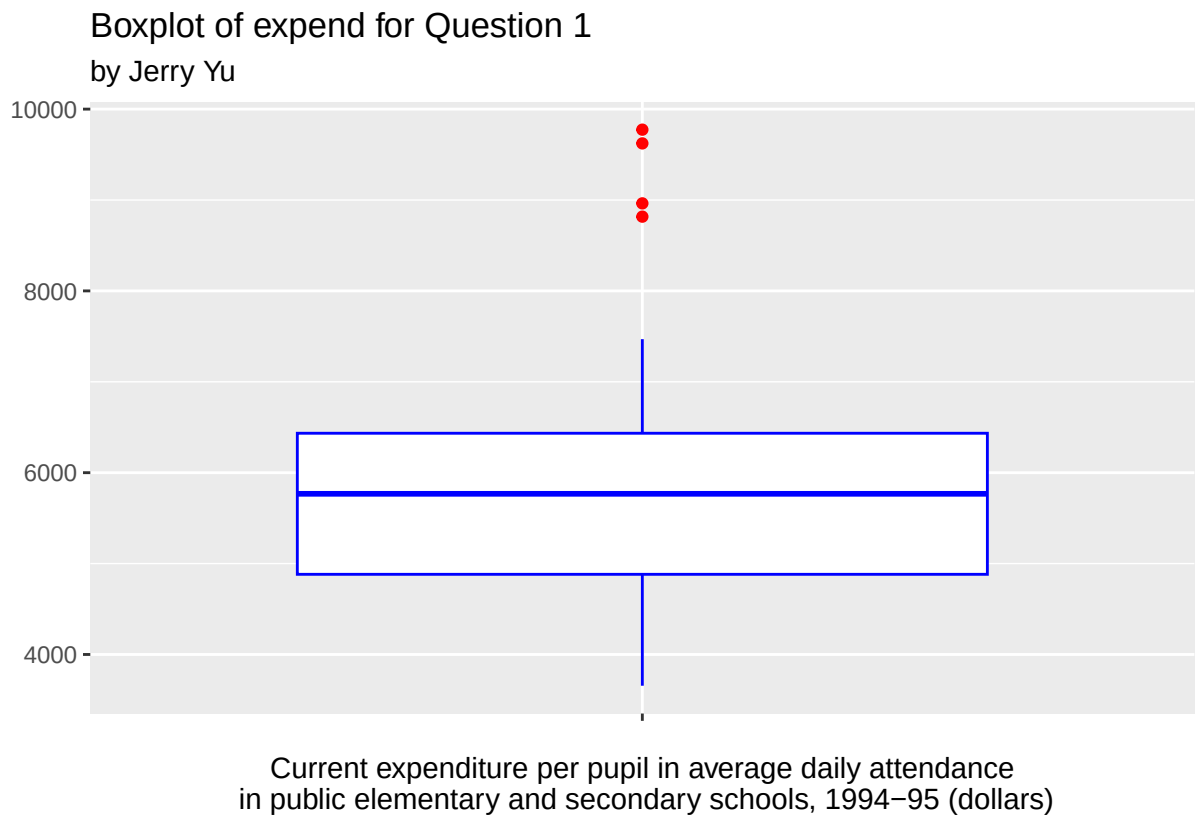
```
##  7 Connecticut   8817  14.4  50045     81    431   477   908
##  8 Delaware      7030  16.6  39076     68    429   468   897
##  9 Florida       5718  19.1  32588     48    420   469   889
## 10 Georgia       5193  16.3  32291     65    406   448   854
## # ... with 40 more rows
```

Then we run Pima and Prostate Style preanalysis.

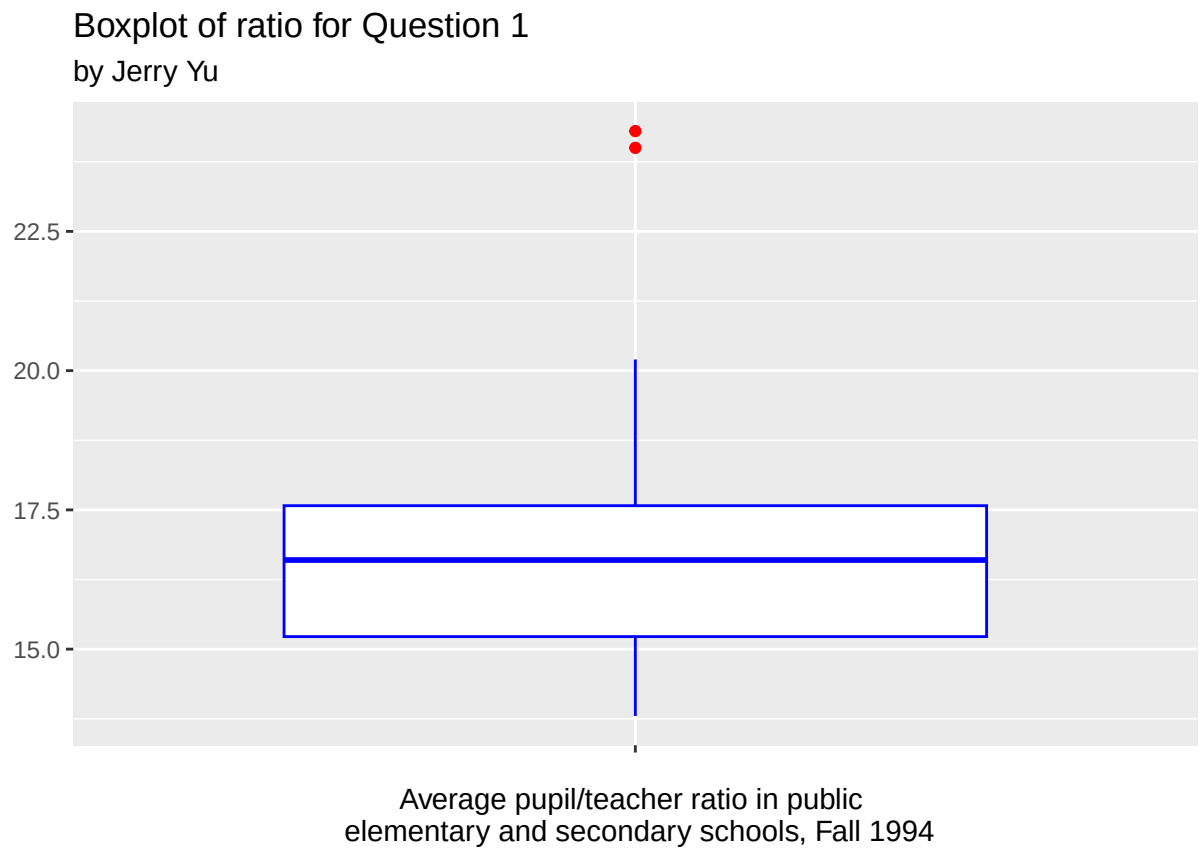
```
summary(sat)
```

```
##     State               expend         ratio           salary
##  Length:50          Min.   :3656   Min.   :13.80   Min.   :25994
##  Class :character   1st Qu.:4882   1st Qu.:15.22   1st Qu.:30978
##  Mode  :character   Median :5768   Median :16.60   Median :33288
##                     Mean   :5905   Mean   :16.86   Mean   :34829
##                     3rd Qu.:6434   3rd Qu.:17.57   3rd Qu.:38546
##                     Max.   :9774   Max.   :24.30   Max.   :50045
##      takers          verbal           math           total
##  Min.   : 4.00   Min.   :401.0   Min.   :443.0   Min.   : 844.0
##  1st Qu.: 9.00   1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :28.00   Median :448.0   Median :497.5   Median : 945.5
##  Mean   :35.24   Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:63.00   3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :81.00   Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

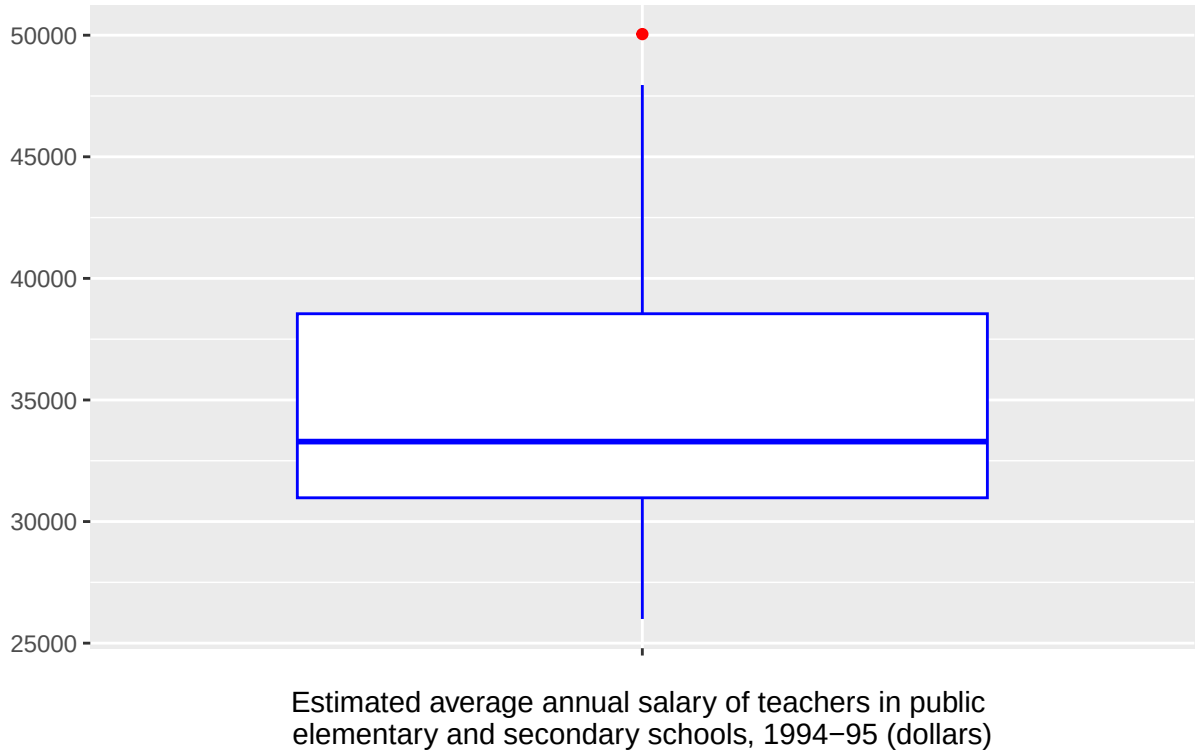
```
#Boxplots

create_boxplot <- function(data, v, xlab, q) {
  ggplot(data = data, aes(x = "", y = !!sym(v))) +
    geom_boxplot(color = "blue", outlier.color = "red") +
    labs(title = paste("Boxplot of", v, "for Question", q),
         subtitle = "by Jerry Yu") +
    xlab(xlab) +
    ylab("")
}

create_boxplot(
  sat,
  "expend",
  "Current expenditure per pupil in average daily attendance\n in public elementary and secondary schoo
",
"1"
)
```



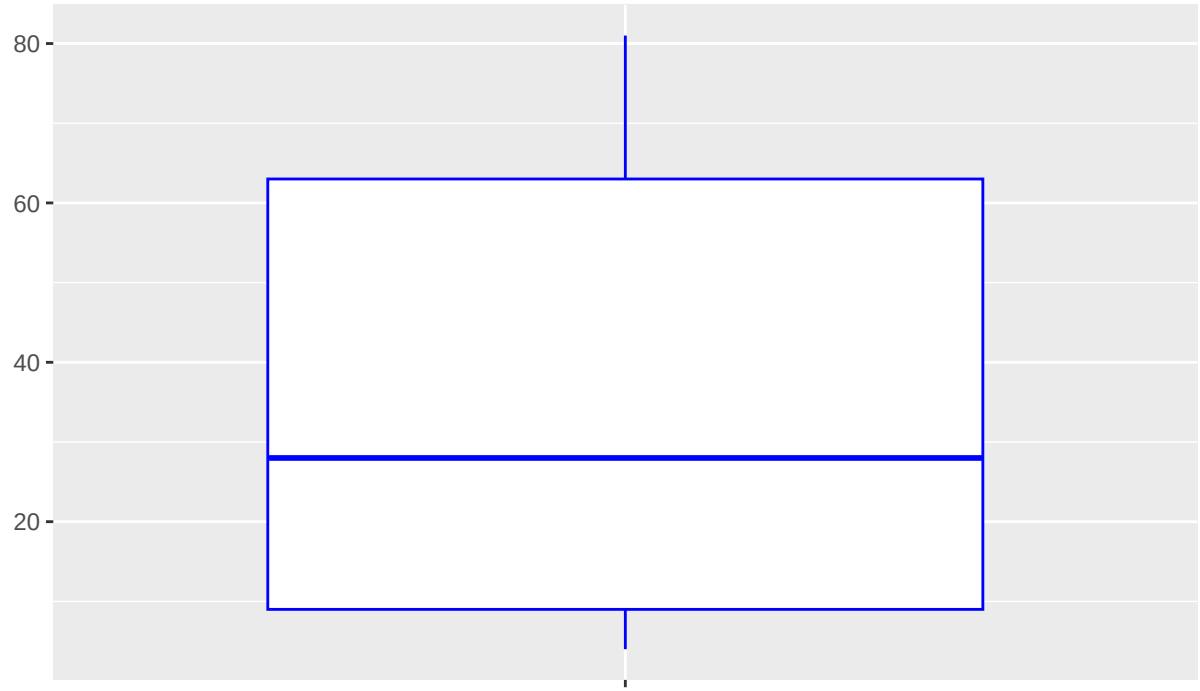

```
create_boxplot(
  sat,
  "ratio",
  "Average pupil/teacher ratio in public \n elementary and secondary schools, Fall 1994",
  "1"
)
```



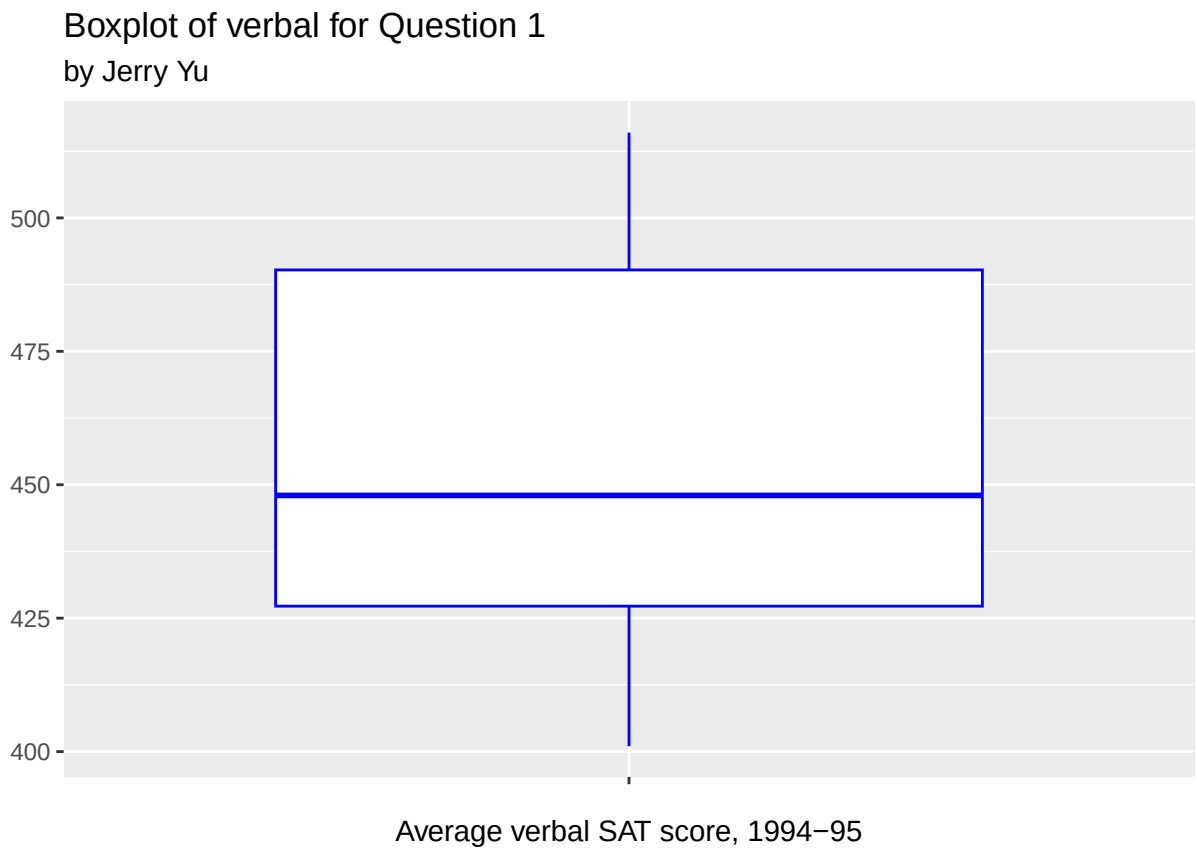

```
create_boxplot(
  sat,
  "salary",
  "Estimated average annual salary of teachers in public \n elementary and secondary schools, 1994-95 (c
  "1"
)
```



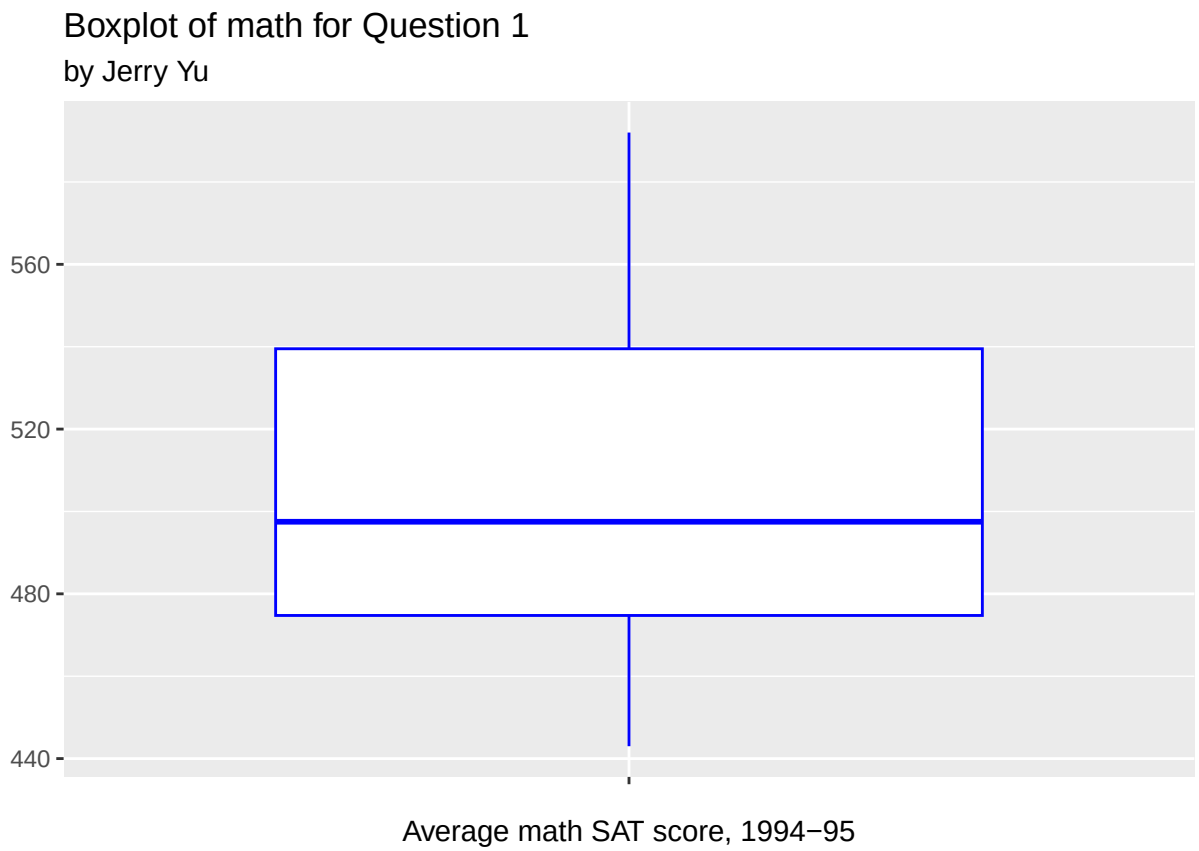

```
create_boxplot(sat,
               "takers",
               "Percentage of all eligible students taking the SAT, 1994-95",
               "1")
```



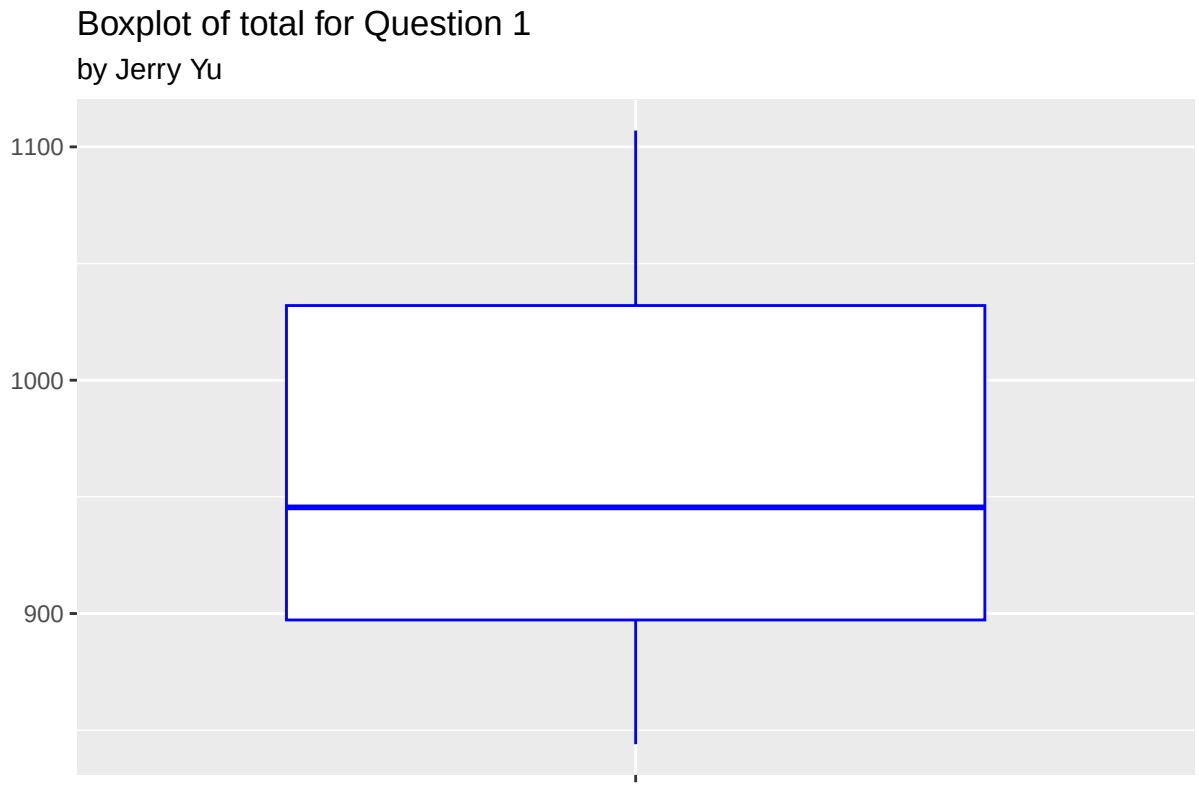

```
create_boxplot(sat, "verbal", "Average verbal SAT score, 1994-95", "1")
```




```
create_boxplot(sat, "math", "Average math SAT score, 1994-95", "1")
```

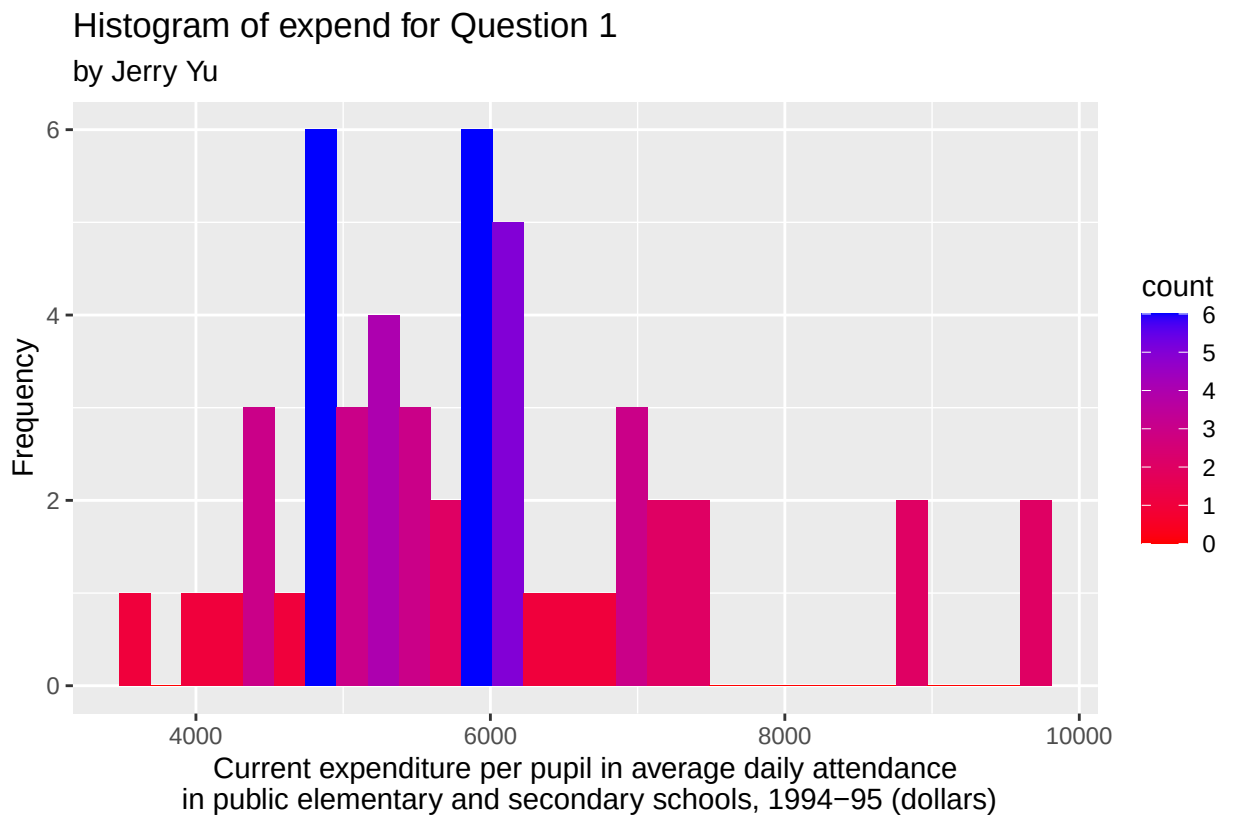



```
create_boxplot(sat, "total", "Average total score on the SAT, 1994-95", "1")
```



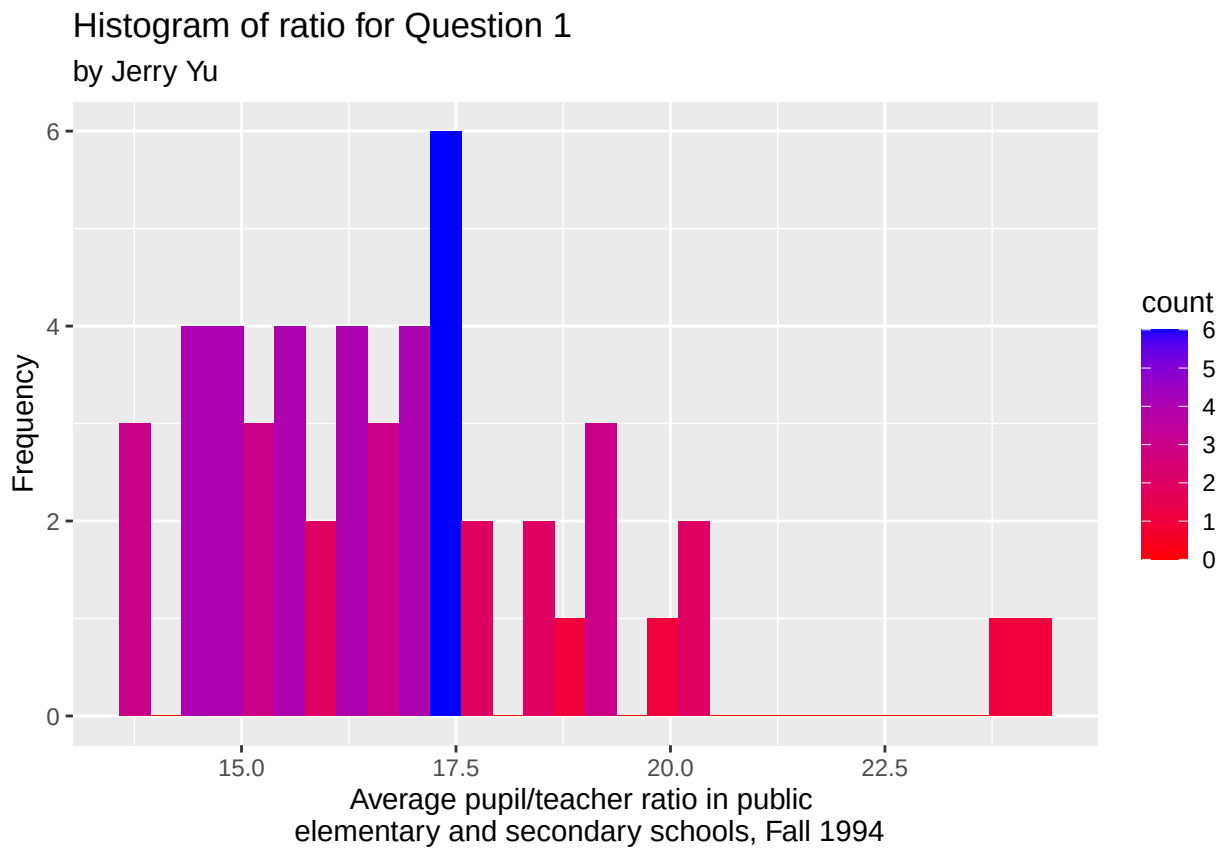

```
# Create Histogram

create_histogram <- function(d, v, xlab, q) {
  ggplot(data = d, aes(x = !!sym(v), fill = ..count..)) +
    geom_histogram() +
    scale_fill_gradient(low = "red", high = "blue") +
    labs(title = paste("Histogram of", v, "for Question", q),
         subtitle = "by Jerry Yu") +
    xlab(xlab) +
    ylab("Frequency")
}

create_histogram(
  sat,
  "expend",
  "Current expenditure per pupil in average daily attendance\n in public elementary and secondary schoo
",
"1"
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



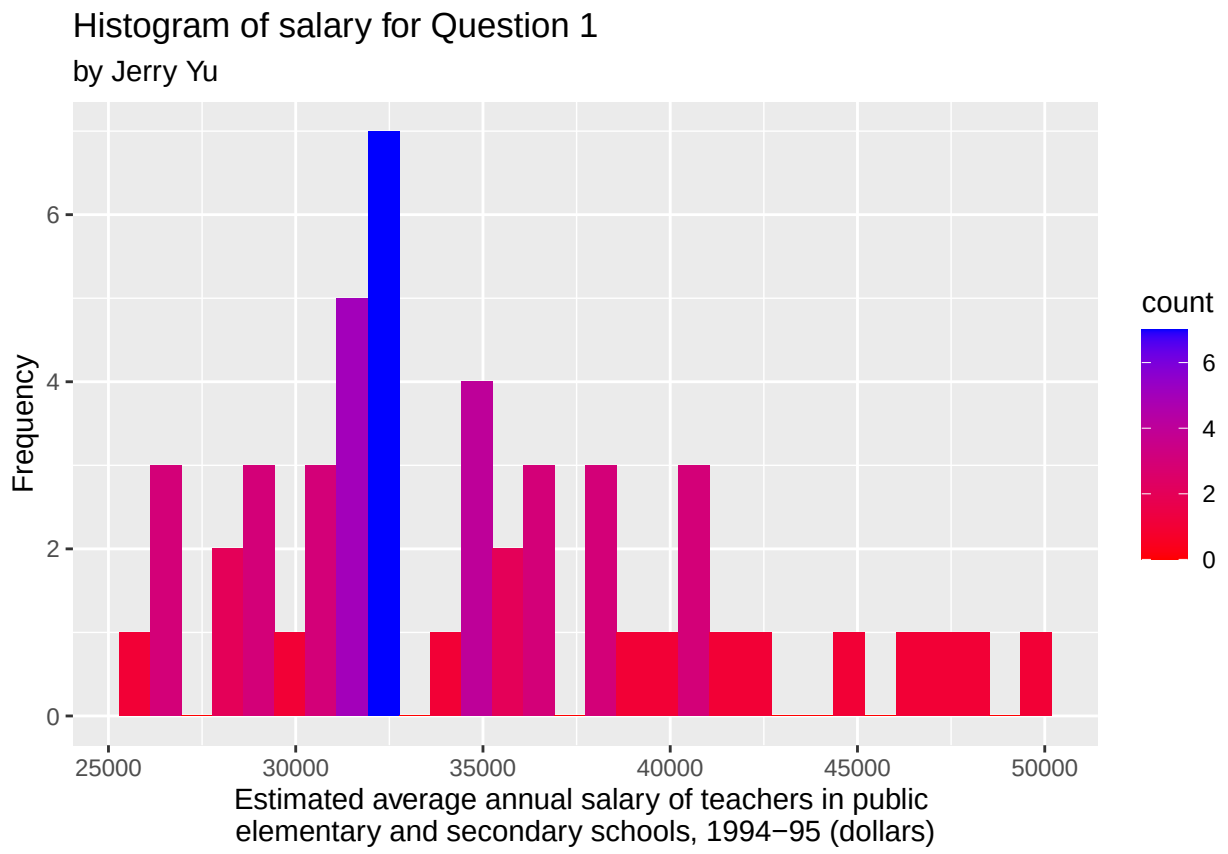

```
create_histogram(
  sat,
  "ratio",
  "Average pupil/teacher ratio in public \n elementary and secondary schools, Fall 1994",
  "1"
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



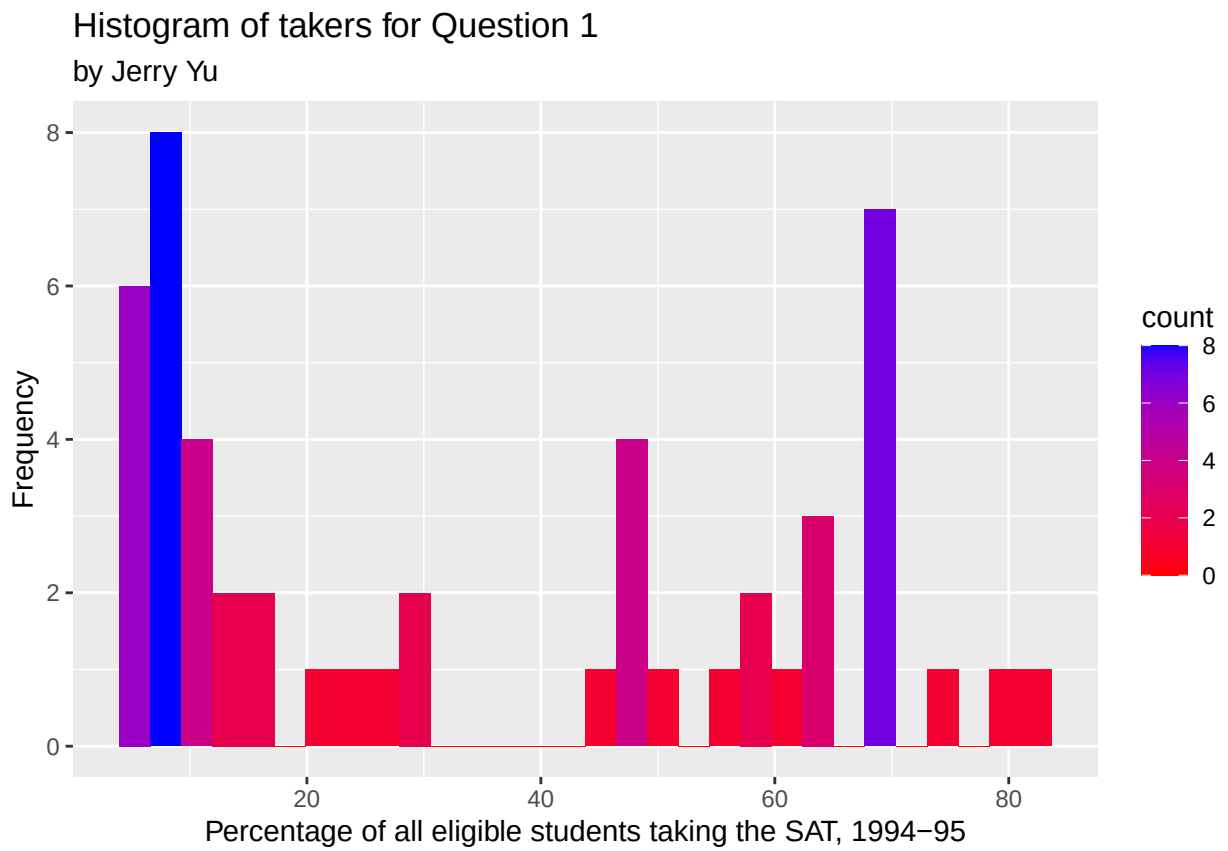

```
create_histogram(
  sat,
  "salary",
  "Estimated average annual salary of teachers in public \n elementary and secondary schools, 1994-95 (c
  "1"
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



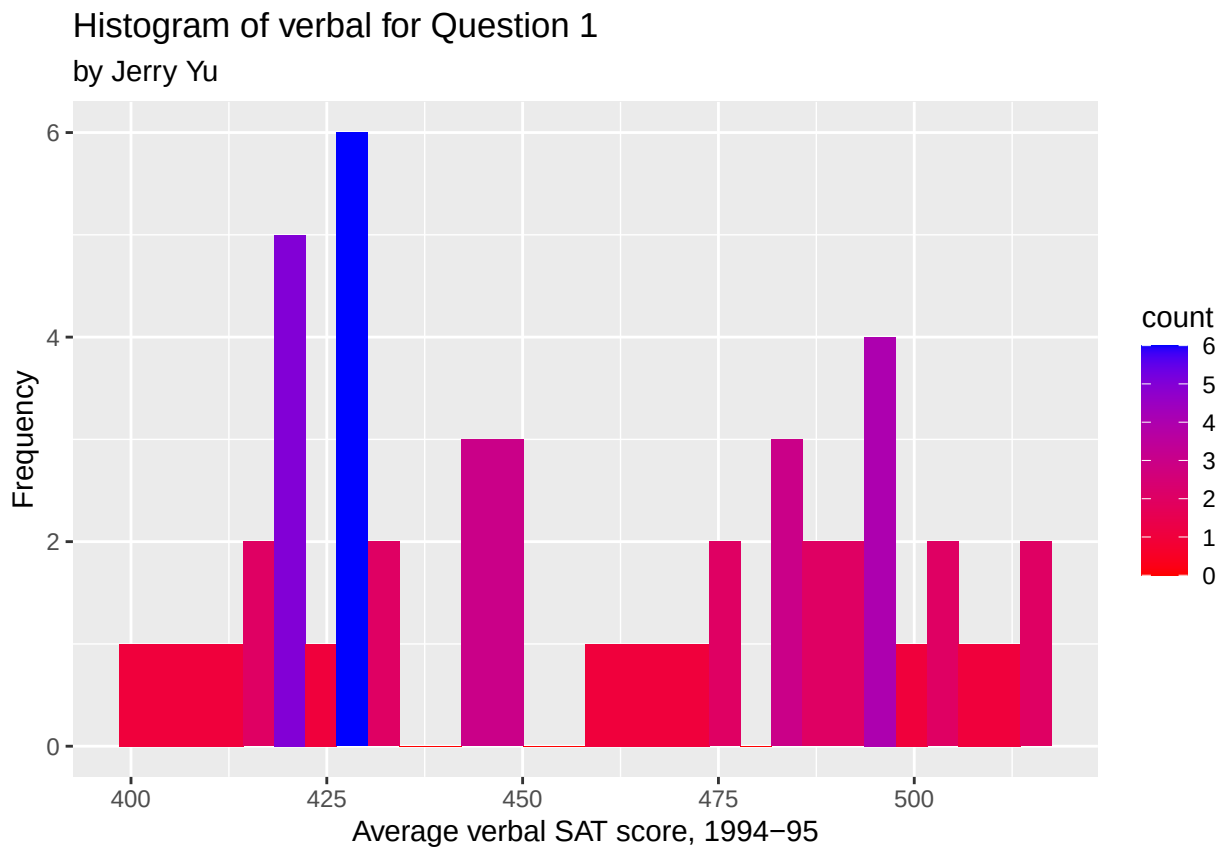

```
create_histogram(sat,
                 "takers",
                 "Percentage of all eligible students taking the SAT, 1994-95",
                 "1")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



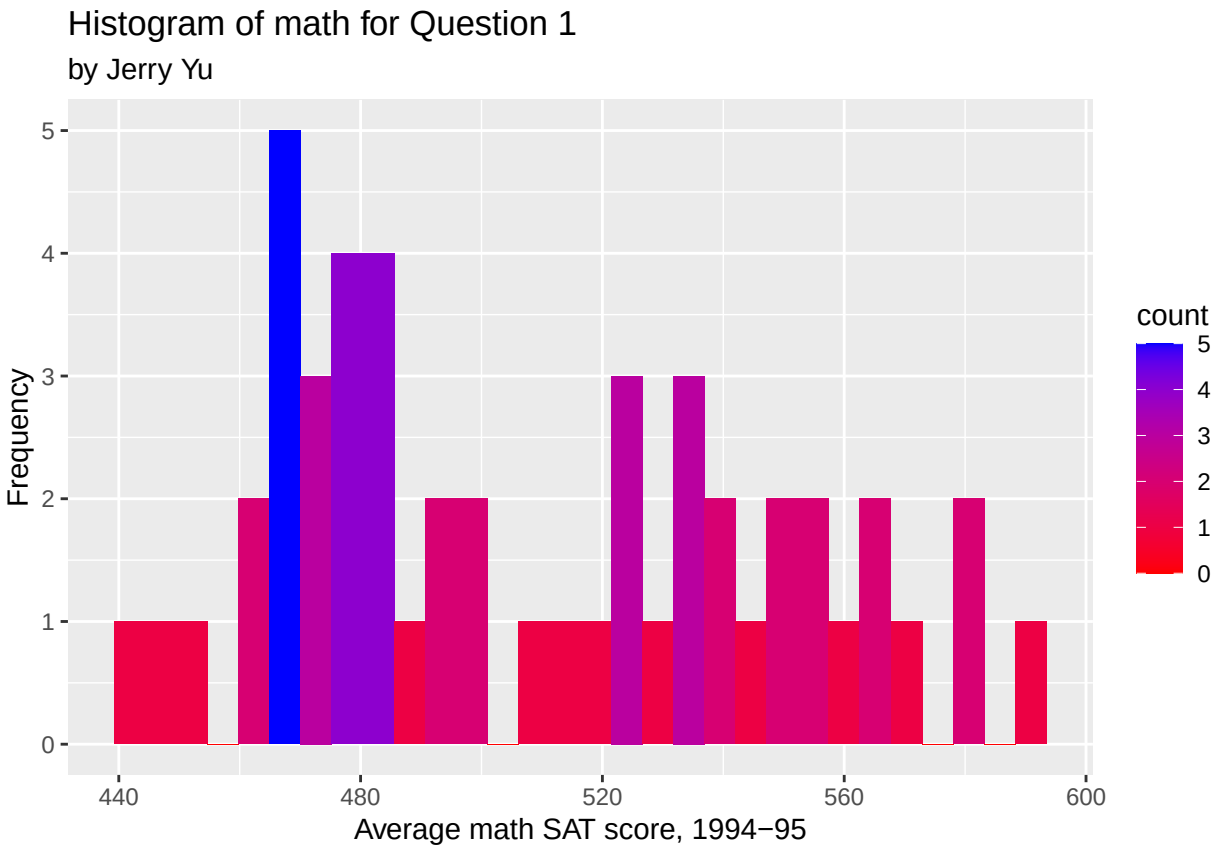

```
create_histogram(sat, "verbal", "Average verbal SAT score, 1994-95", "1")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



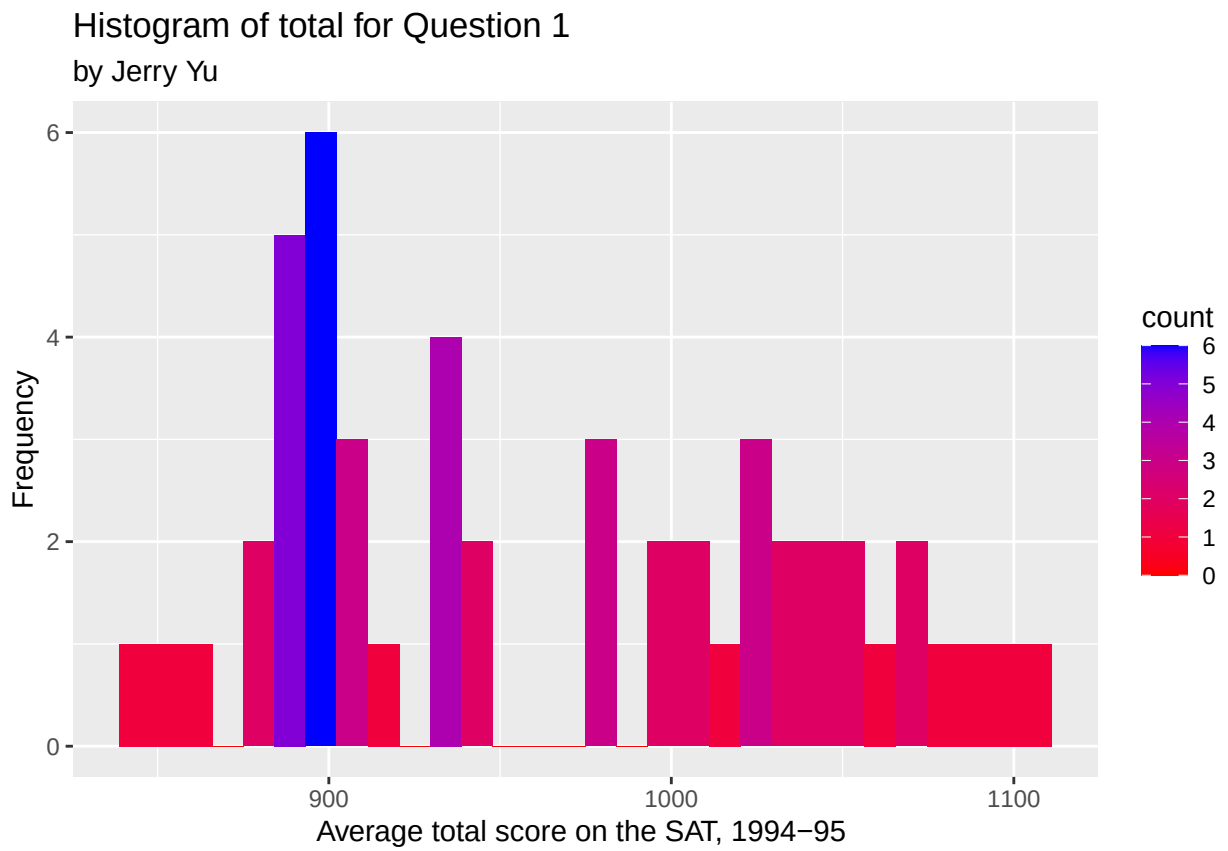

```
create_histogram(sat, "math", "Average math SAT score, 1994-95", "1")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
create_histogram(sat, "total", "Average total score on the SAT, 1994-95 ", "1")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the preliminary data analysis, we see that almost all the distributions are right skewed. However, the degree of skew is more or less extreme depending on the variable, with some variables like expend having many more outliers than variables like math. The similar distributions suggest that the money based variables (like expend and ratio) might be positively correlated with the performance based variables (like math and total). The only exception to this general pattern is takers, which seems bimodal. This suggests that takers might be a confounding variable.

Question 2

> The dataset press.txt for this assignment comes from some research into the production of textiles Do some short numerical and graphical summaries of the data, commenting on any features that you find interesting.

```
press <- read.table("http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/data/wrinkle.txt",header=TRU
```

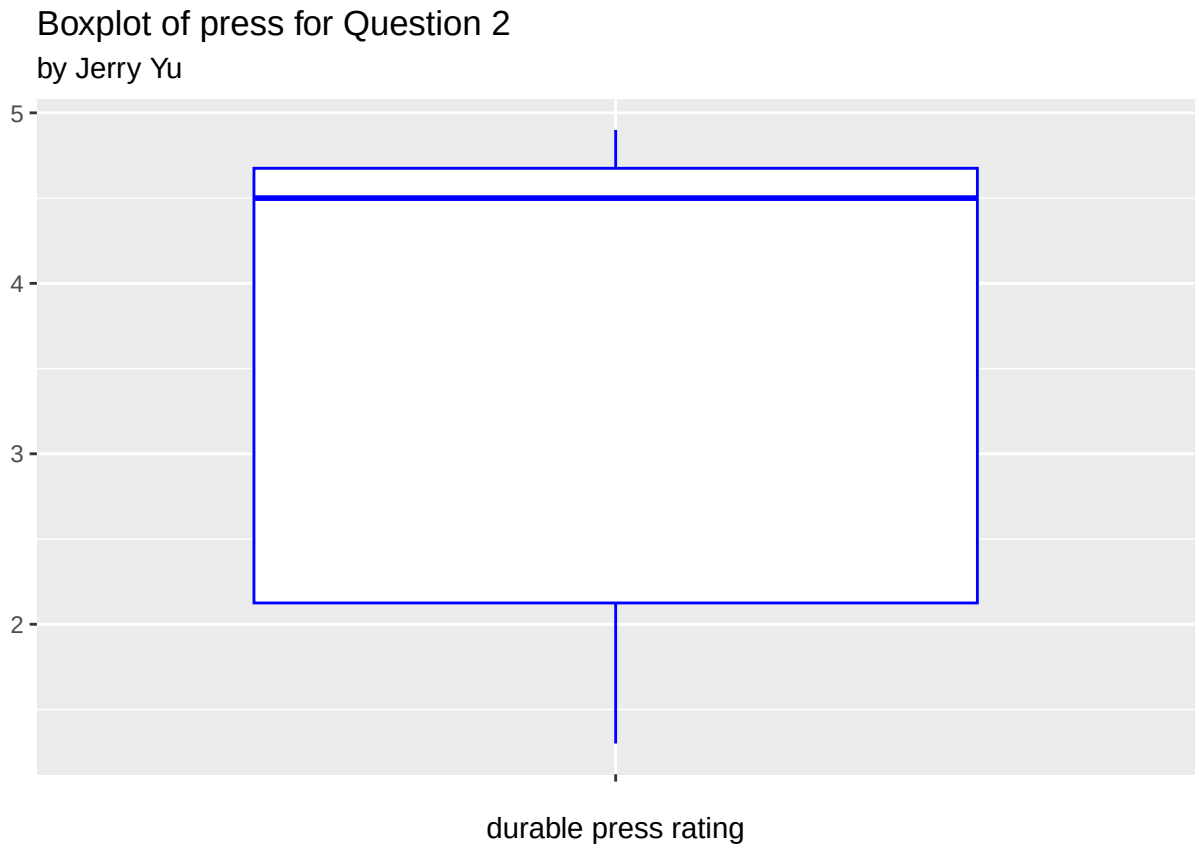
Preanalysis Chunk:

```
summary(press)
```

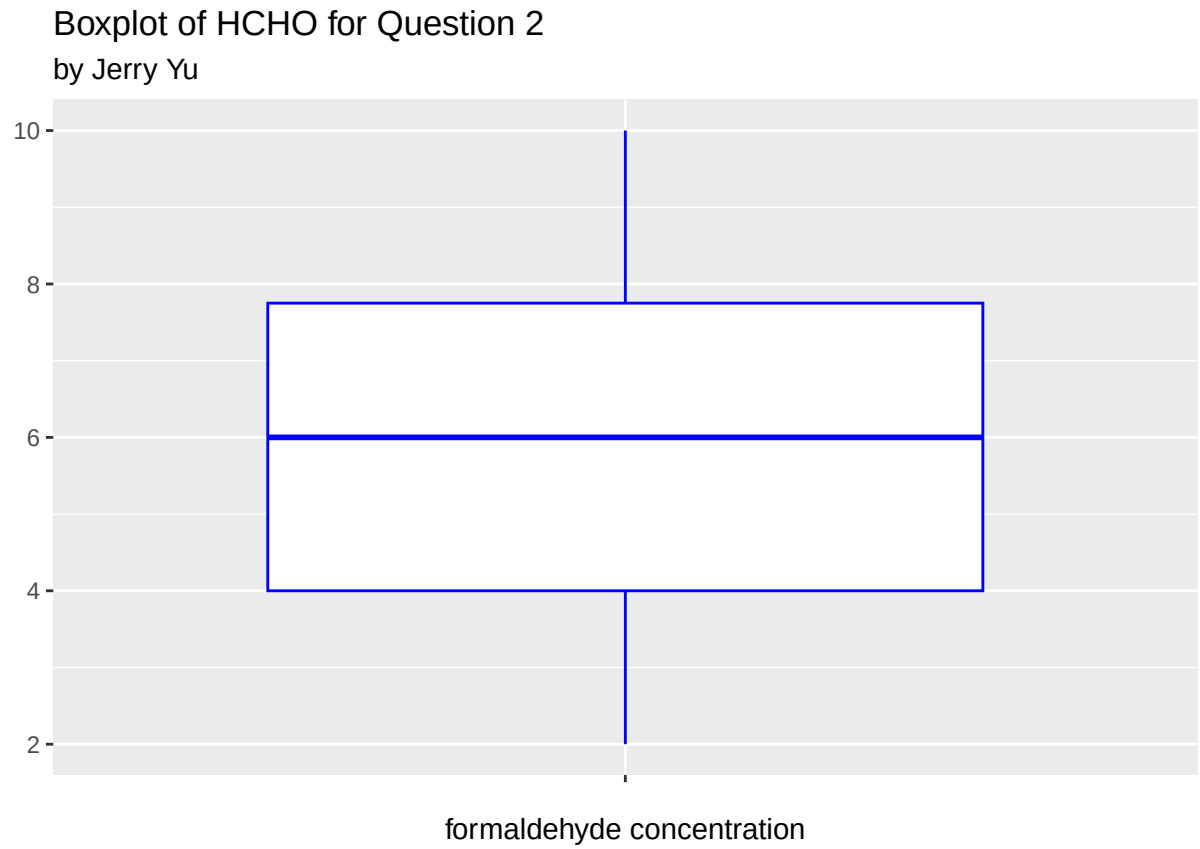
```
##      press            HCHO          catalyst         temp
##  Min.   :1.300   Min.   : 2.000   Min.   : 1.0   Min.   :100.0
##  1st Qu.:2.125   1st Qu.: 4.000   1st Qu.: 4.0   1st Qu.:120.0
##  Median :4.500   Median : 6.000   Median : 7.0   Median :140.0
```

```
##  Mean   :3.560   Mean   : 6.067   Mean   : 6.8   Mean   :142.7
##  3rd Qu.:4.675   3rd Qu.: 7.750   3rd Qu.:10.0   3rd Qu.:180.0
##  Max.   :4.900   Max.   :10.000   Max.   :13.0   Max.   :180.0
##       time
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :3.000
##  Mean   :3.933
##  3rd Qu.:7.000
##  Max.   :7.000
```

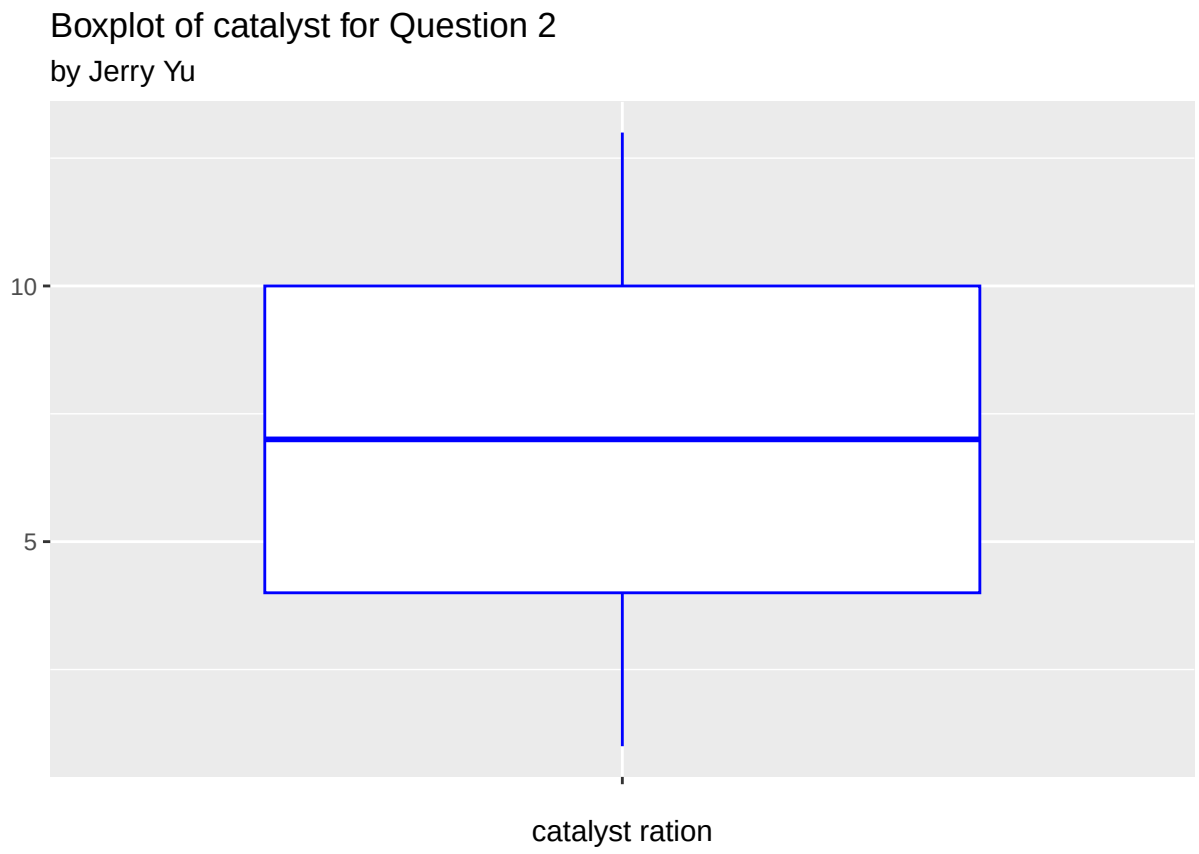
```
#boxplot
create_boxplot(press, "press",
               "durable press rating", "2")
```



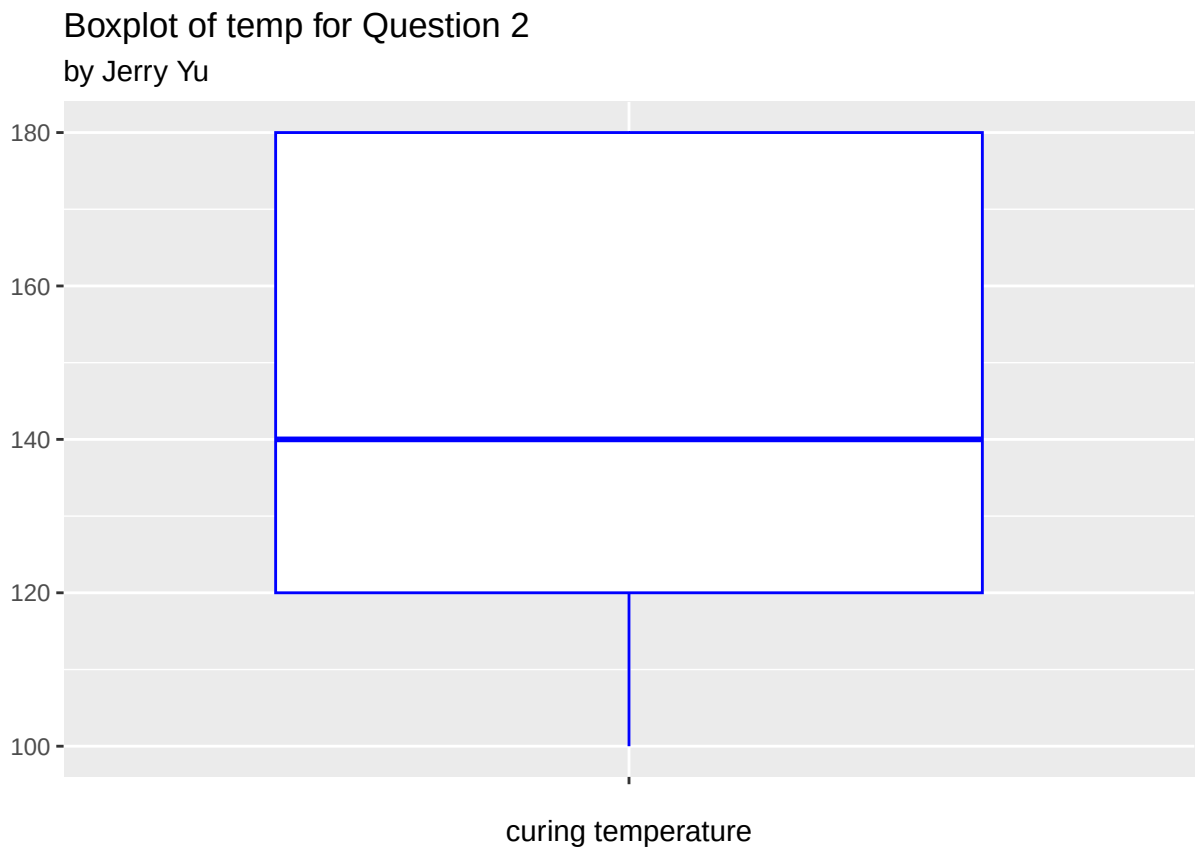

```
create_boxplot(press, "HCHO",
               "formaldehyde concentration", "2")
```



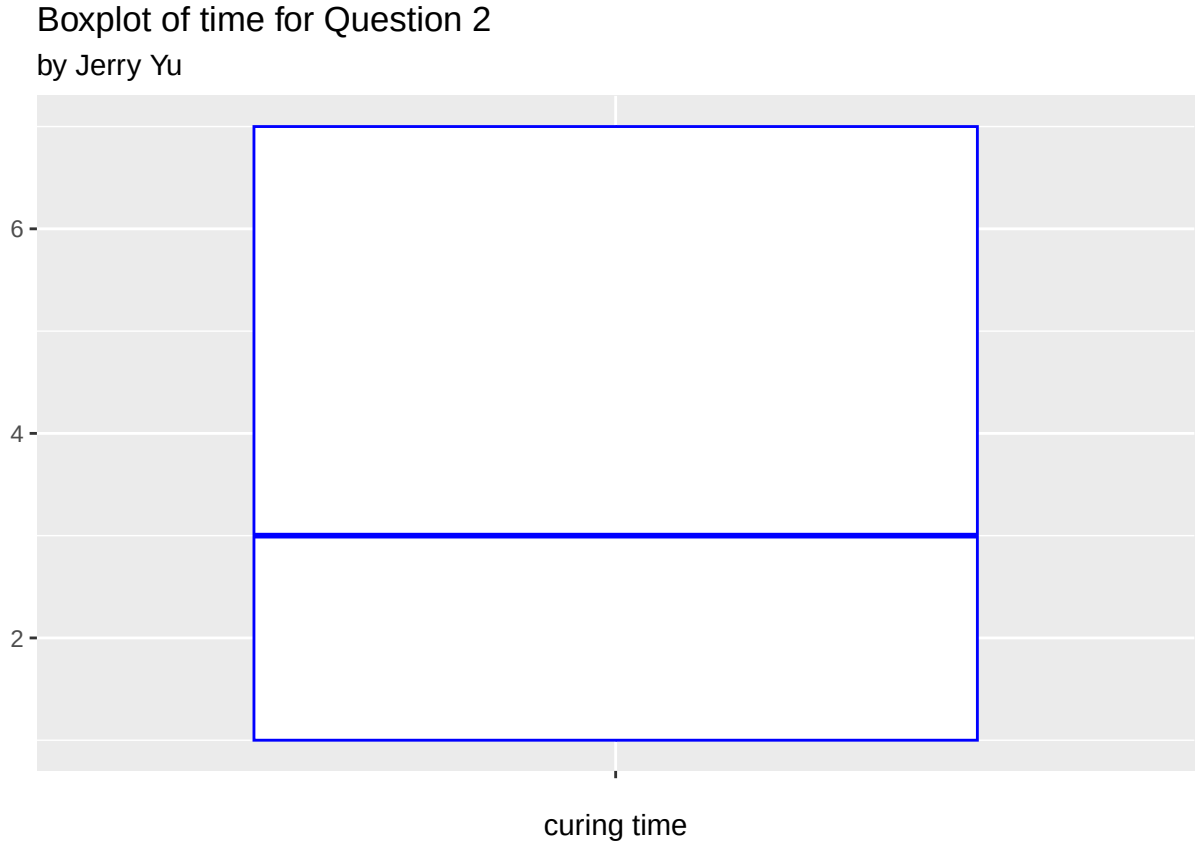

```
create_boxplot(press, "catalyst",
               "catalyst ration", "2")
```




```
create_boxplot(press, "temp",
               "curing temperature", "2")
```

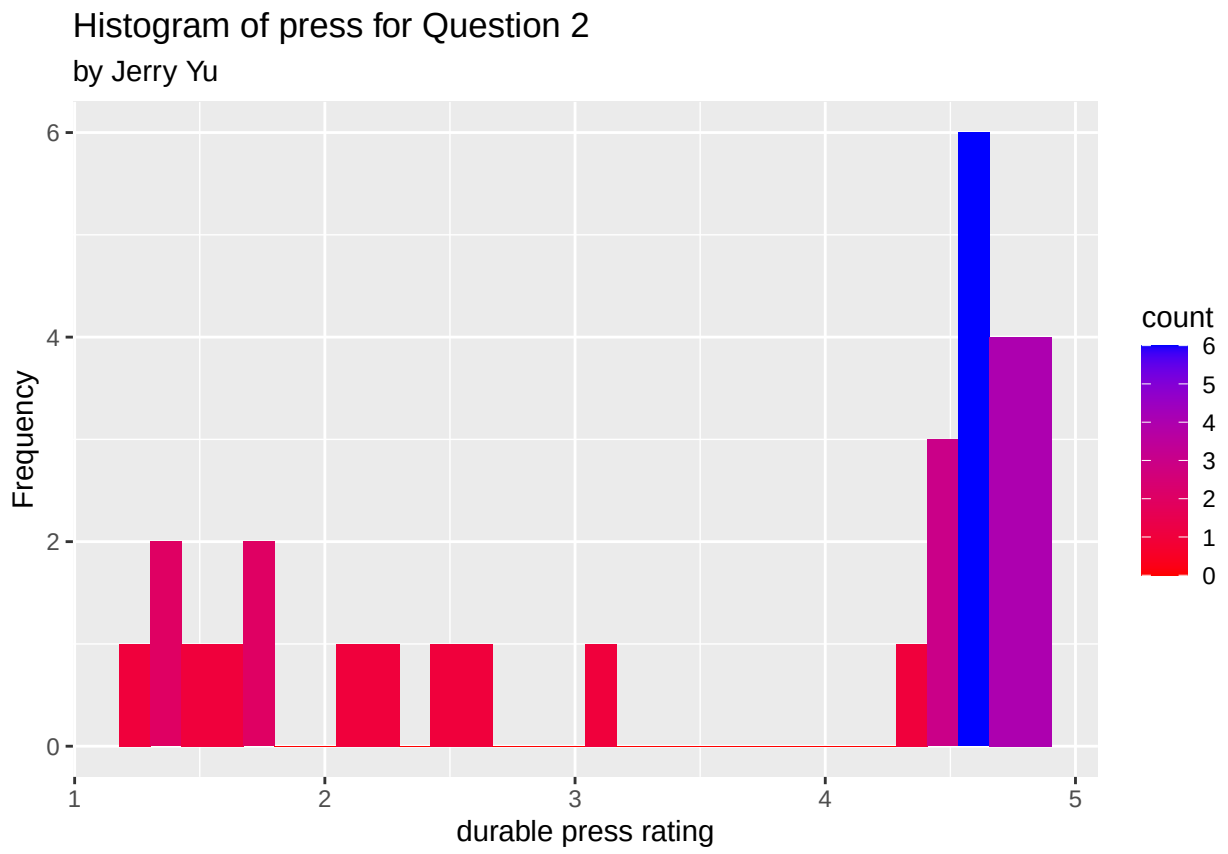



```
create_boxplot(press, "time",
               "curing time", "2")
```



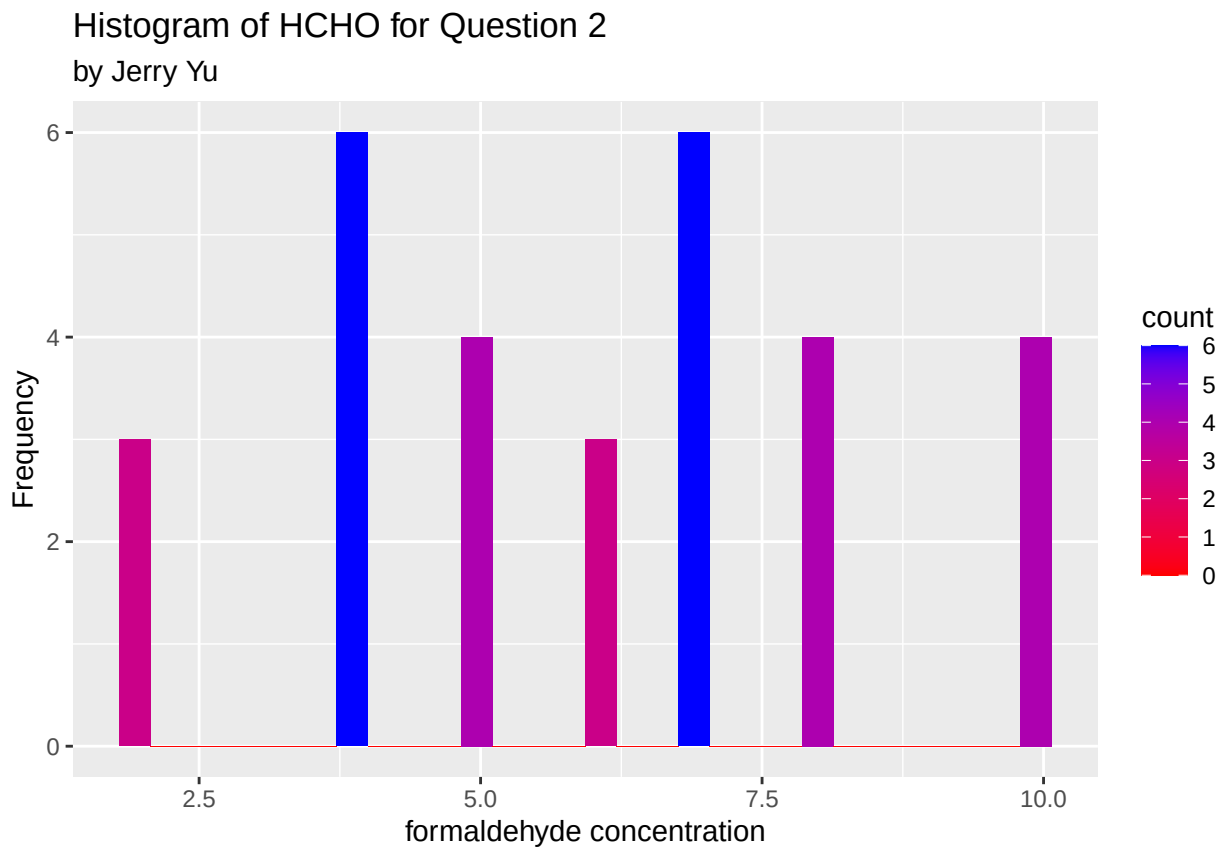

```
#histogram
create_histogram(press, "press",
                 "durable press rating", "2")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



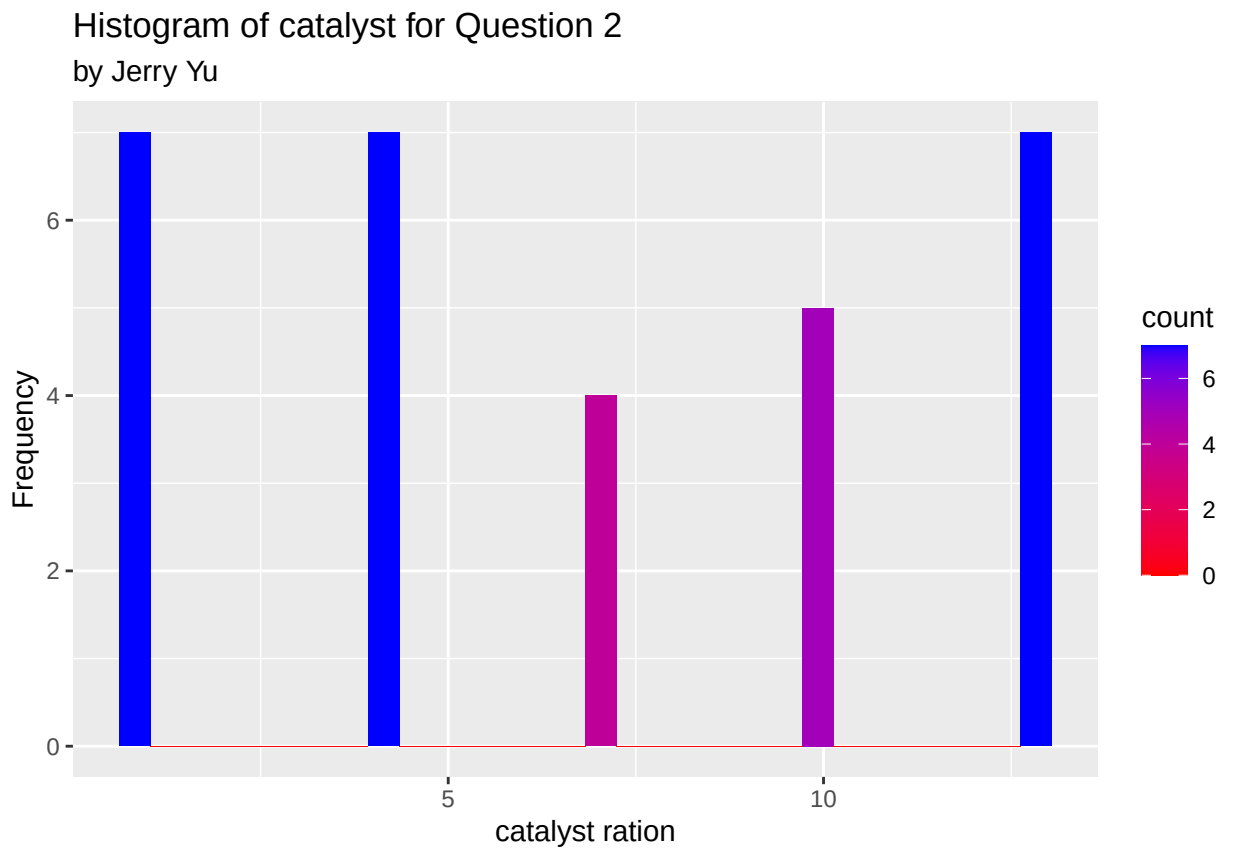

```
create_histogram(press, "HCHO",
                 "formaldehyde concentration", "2")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



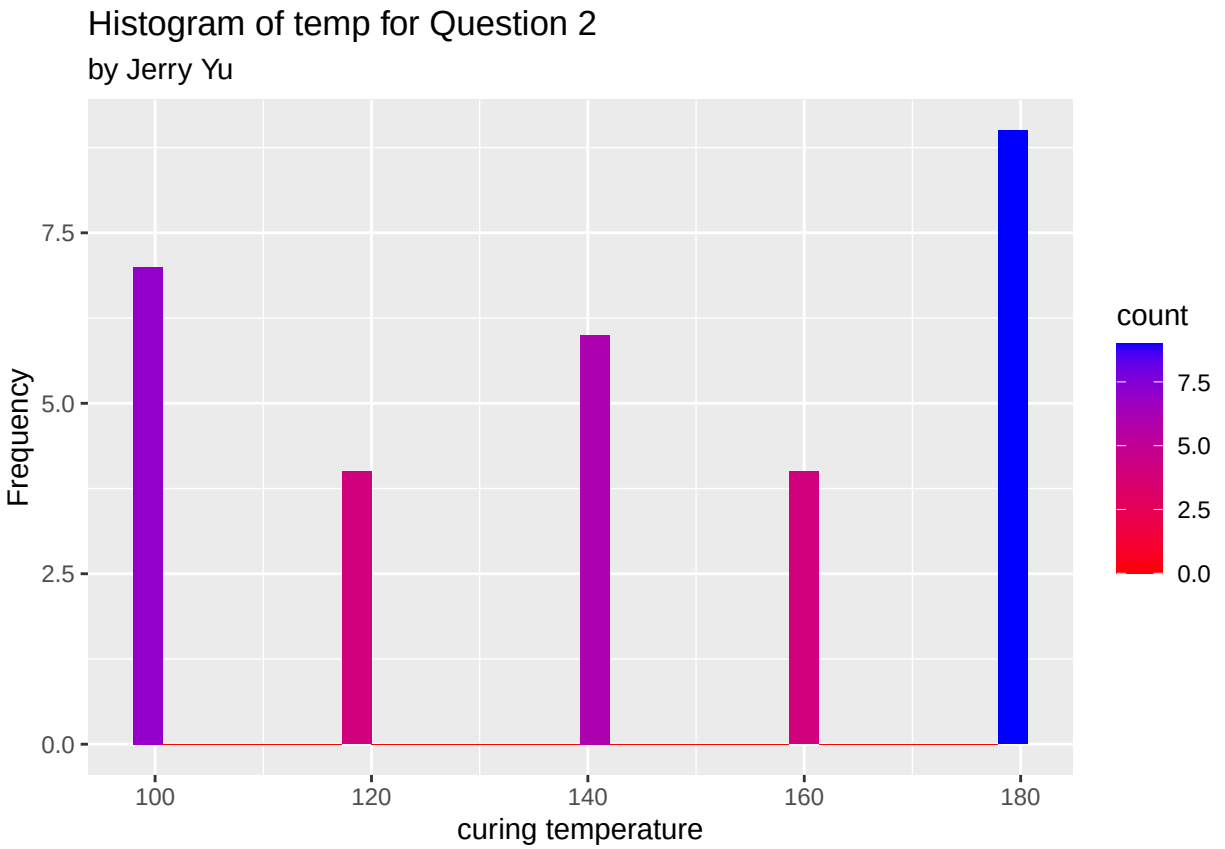

```
create_histogram(press, "catalyst",
                 "catalyst ration", "2")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



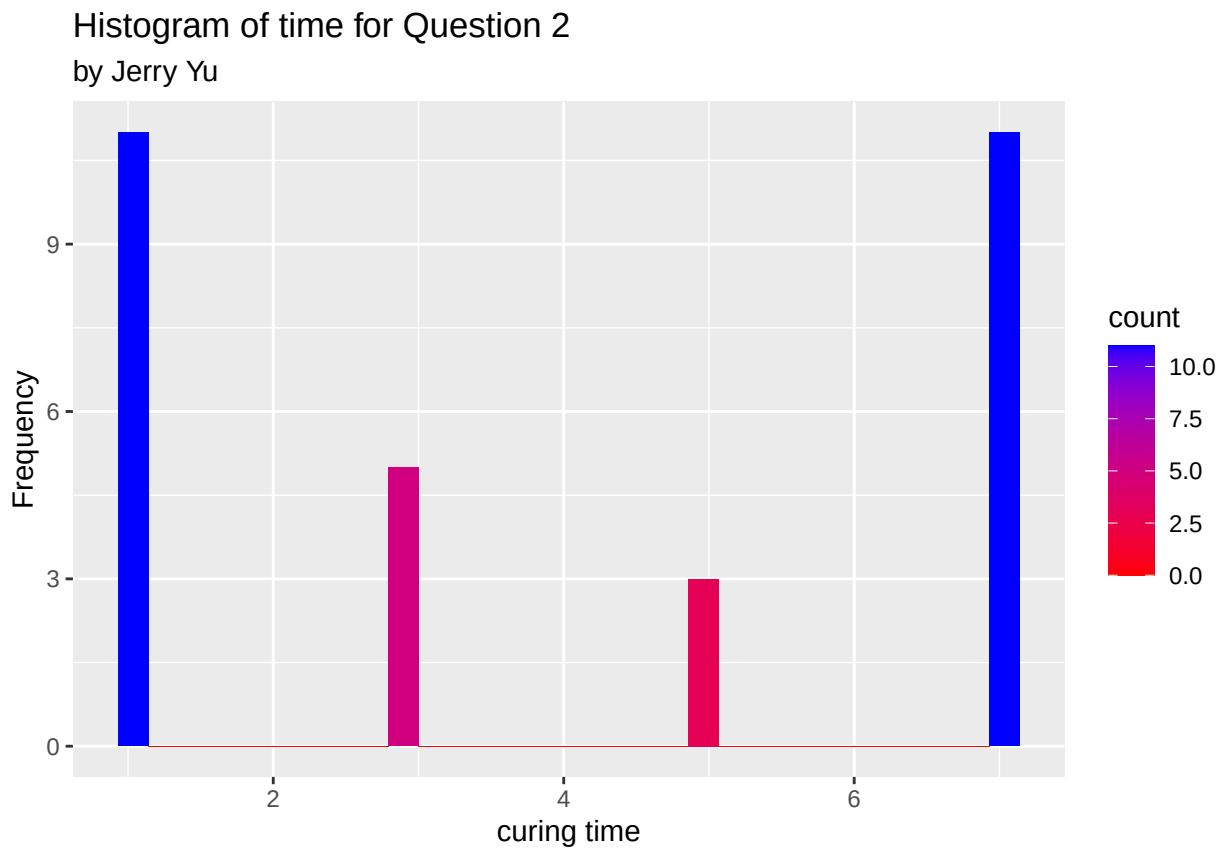

```
create_histogram(press, "temp",
                 "curing temperature", "2")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
create_histogram(press, "time",
                 "curing time", "2")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From our preanalysis, we see that our 5 variables generally fall into 3 distributions. Press is left skewed, HCHO is relatively symmetric, and catalyst, temp, and time are all strongly bimodal. Additionally, these data points seem more uniform, with no outliers for all 5 variables and Q3 = max temp and time.