Homework 3 Solutions

1. Load the Camden Boroughs dataset.

```
setwd("/Users/Daniel/Dropbox/Teaching/CourseR/")
d <- read.csv("./data/CamdenBoroughs.csv", na.strings = c("Not applicable",
       "Unknown", "999", "NA", ""))
```

2. Create a new variable, `frl`, that is a numeric version of `Percentage.Claiming.Free.School.Meals`.

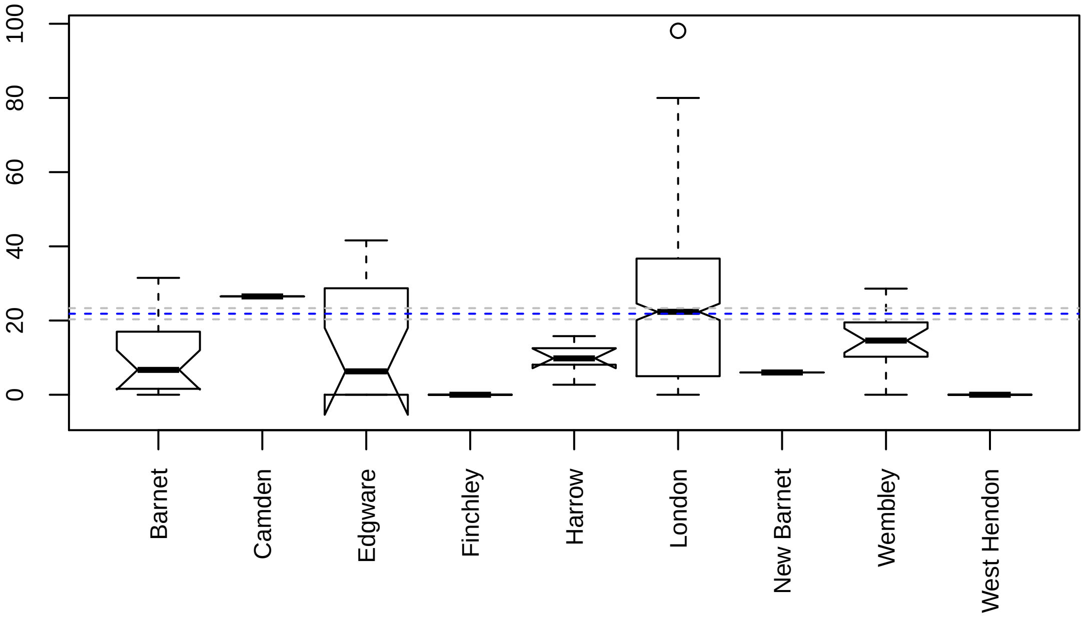
```
d$frl <- as.numeric(sub("%", "", d$Percentage.Claiming.Free.School.Meals))
```

3. Create a notched boxplot of `frl` by `Town`. Add a horizontal blue dashed line to the plot displaying the grand mean, and gray horizontal lines above and below the grand mean to display the 95% confidence interval around the mean.

```
se <- function(x) sqrt(var(x, na.rm = TRUE)/length(na.omit(x)))

par(las = 3, cex.axis = 0.75)

plot(d$Town, d$frl, notch = TRUE)
abline(h = mean(d$frl, na.rm = TRUE), col = "blue", lty = 2)
abline(h = mean(d$frl, na.rm = TRUE) + 1.96*se(d$frl),
    col = "gray", lty = 2)
abline(h = mean(d$frl, na.rm = TRUE) - 1.96*se(d$frl),
    col = "gray", lty = 2)
```



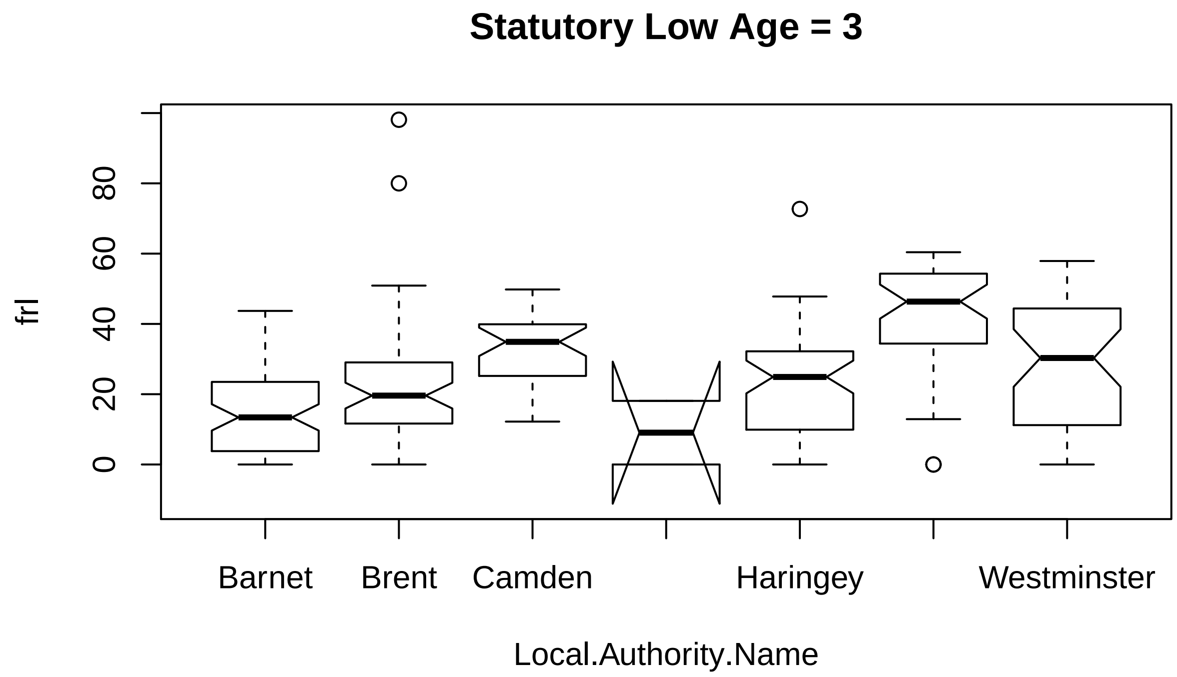
4. Briefly interpret the plot.

London pretty clearly seems to dominate the data, given that the grand mean aligns almost perfectly with the median in London, and the median in all other towns is below the grand mean, with the exception of Camden. Indeed, 593 of the 664 observations, or 89% are from London, while 4 of the towns had only one observation. However, of the groups with more than one observation, the proportion of students eligible for free or reduced price lunch was reliably lower than the grand mean, as the upper-bound of the 95% confidence interval around the median for each group was below the grand mean. Outside of London, Edgware had the most variability.

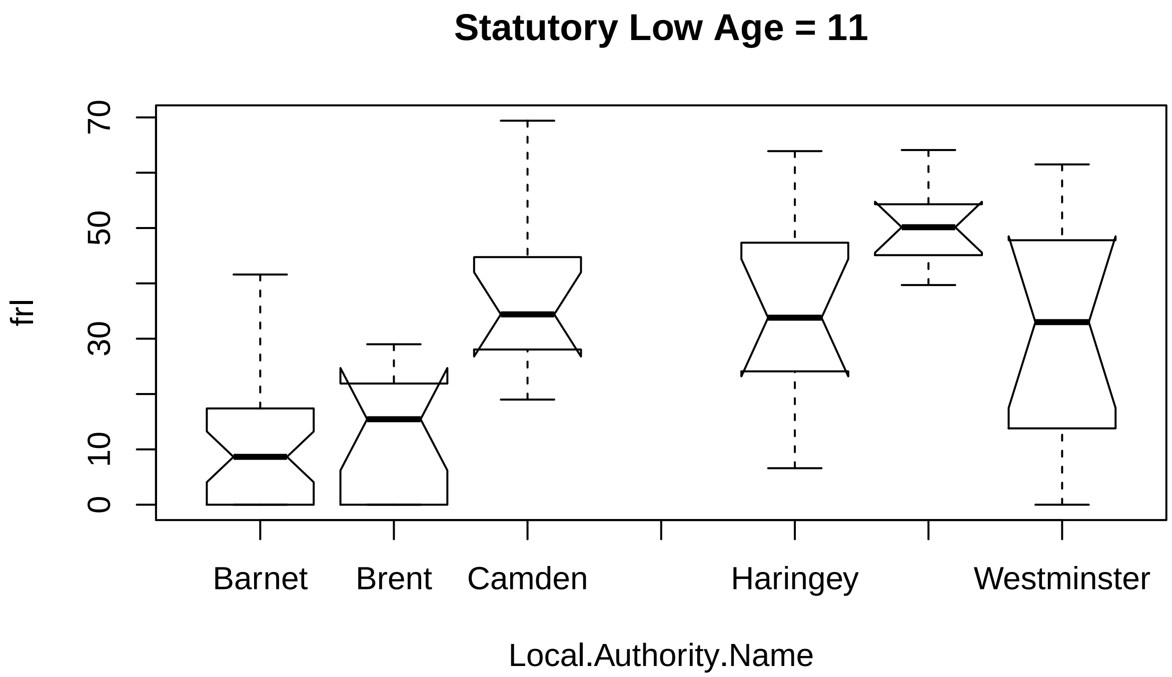
5. Transform the data frame into a list of data frames separated by `Statutory.Low.Age`. The list should be of length 19. Produce a boxplot for `frl` by `Local.Authority.Name` for only schools in which the `Statutory.Low.Age` is 3. Use the list to produce the same plot for schools in which the `Statutory.Low.Age` is 11.

```r
l <- split(d, d$Statutory.Low.Age)

plot(frl ~ Local.Authority.Name,
    data = l[["3"]],
    notch = TRUE,
    main = "Statutory Low Age = 3")
```

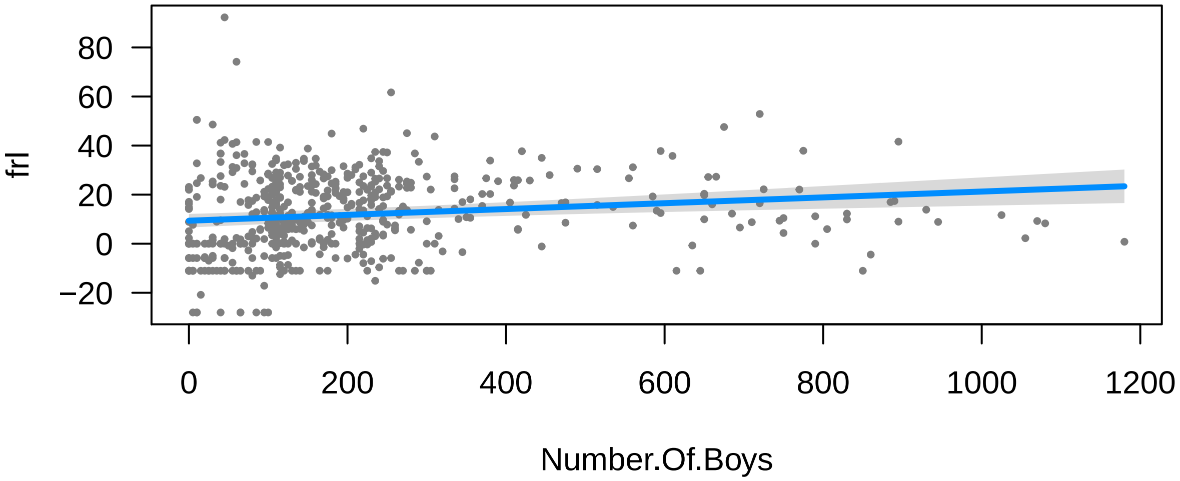


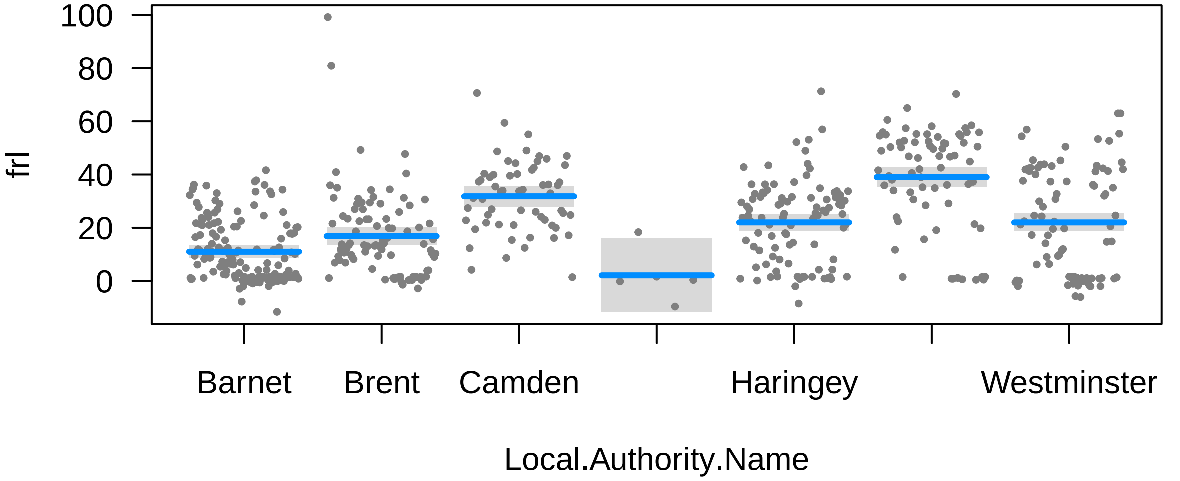
```r
plot(frl ~ Local.Authority.Name,
    data = l[["11"]],
    notch = TRUE,
    main = "Statutory Low Age = 11")
```



6. Fit a multiple regression model for schools in which the Statutory Low Age is 3, with `Number.Of.Boys` and `Local.Authority.Name` are modeled as predictors of `frl`. Compute predictor-residual plots for the model (*hint: use a package*).

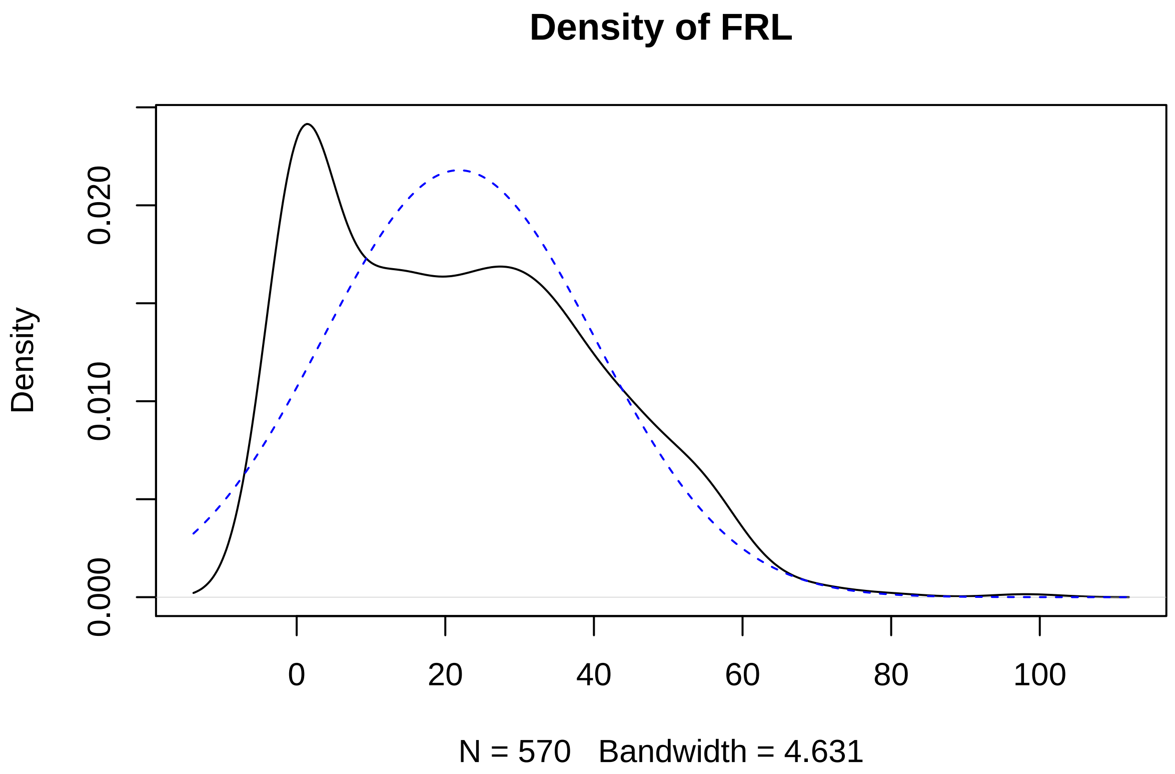
```
mod <- lm(frl ~ Number.Of.Boys + Local.Authority.Name, data = d)
library(visreg)
par(mfrow = c(2,1))
visreg(mod)
```





7. Plot the density of `frl`. Overlay a plot of the likelihood, had the data been generated by a normal distribution with a mean and standard deviation equal to the sample mean and standard deviation.

```r
dens <- density(d$frl, na.rm = TRUE)
plot(dens, main = "Density of FRL")
lines(dens$x,
      dnorm(dens$x, mean(d$frl, na.rm = TRUE), sd(d$frl, na.rm = TRUE)),
      col = "blue",
      lty = 2)
```



Do the data appear to have been generated by such a distribution? Why or why not?

It’s difficult to say for sure if the data were generated by the specified normal distribution, but it appears somewhat unlikely. The empirical density function has a spike around 0 that is not present in the normal distribution, and the peak of the normal is well above the empirical density function.