Likes in 2 variables
Problem 2.10

2.1 $\text{mean}(\{kx\}) = \frac{\sum_{i=1}^{N} kx_i}{N} = k \cdot \frac{\sum_{i=1}^{N} x_i}{N} = k\,\text{mean}(\{x\})$

b/c k is constant

2.2 $\text{mean}(\{x+c\}) = \frac{\sum_{i=1}^{N} x_i + c}{N} = \frac{N \cdot c + \sum_{i=1}^{N} x_i}{N} = c + \frac{\sum_{i=1}^{N} x_i}{N} = c + \text{mean}(\{x\})$

b/c C is constant

2.3 $\sum_{i=1}^{N} (x_i - \text{mean}(\{x\})) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \text{mean}(\{x\}) = \sum_{i=1}^{N} x_i - N \cdot \text{mean}(\{x\}) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0$

2.2

2.4 $\text{std}(\{x+c\}) = \sqrt{\frac{\sum_{i=1}^{N} (x_i + c - \text{mean}(\{x+c\}))^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (x_i + c - (\text{mean}(x) + c))^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \text{mean}(\{x\}))^2}{N}}$

Mean is constant with respect to i

$= \text{std}(\{x\})$ ✓

2.5 $\text{std}(\{kx\}) = \sqrt{\frac{\sum_{i=1}^{N} (kx_i - \text{mean}(\{kx\}))^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (kx_i - k\,\text{mean}(\{x\}))^2}{N}}$

$= \sqrt{\frac{k^2 \sum_{i=1}^{N} (x_i - \text{mean}(\{x\}))^2}{N}} = k\sqrt{\frac{\sum_{i=1}^{N} (x_i - \text{mean}(\{x\}))^2}{N}}$

$= k \cdot \text{std}(\{x\})$

2.6 suppose x is a sorted set. Let $x_i = \text{median}(\{x\})$

adding c to each element will not change the order of its elements ($y > z \rightarrow y + c > z + c$), so the new median is $x_i + c$

N is odd / N is even

$\{x\} = \{x_1, x_2 \ldots x_i, x_{i+1} \ldots, x_N\}$ $\text{med}(\{x\}) = x_i$ or $\frac{x_i + x_{i+1}}{2}$

$\{x+c\} = \{x_1+c, x_2+c \ldots x_i+c, x_{i+1}+c, \ldots x_N+c\}$ $\text{med}(\{x+c\}) = x_i + c$ or $\frac{x_i + x_{i+1} + 2c}{2}$

$= \text{med}(\{x\}) + c$

2.7 $\{kx\} = \{kx_1, \ldots kx_i, kx_{i+1}, \ldots kx_N\}$ $\text{med}(\{kx\}) = kx_i$ or $k\left(\frac{x_i + x_{i+1}}{2}\right) = k\,\text{med}(x)$

if $k<0$, order is reversed, but for odd N, kx_i is still median; so the proof stands

For even N, $k<0$, order is reversed, so positions of x_i and x_{i+1} are switched, but $k(x_i + x_{i+1}) = (x_{i+1} + x_i)k$, so the proof also stands

2.8 $\text{iqr}(\{x\}) = (x_i) - (x_j)$ | where x_i is 75th or percentile and x_j is 25th

$\text{iqr}(\{x+c\}) = (x_i + c) - (x_j + c) = x_i - x_j = \text{iqr}(\{x\})$

because order is preserved $a > b \rightarrow a + c > b + c$

2.9 iqr({x}) x_i ...

for k>0 iqr({kx}), order is preserved, so iqr({kx}) = $kx_i - kx_j$ = $k(x_i - x_j)$ = k iqr({x})

for k<0 iqr({kx}), order reversed, but i and j are [illegible] swapped [illegible] size of magnitude [illegible] preserved, so iqr({kx}) = $kx_j - kx_i$ = k iqr({x})

k = 0, iqr 0 = 0 iqr({x})

2.10 No. The graph has a pretty clear exponential trend with respect to time, so a constant function like mean would be a poor representation of the data.

2.11 a. I'm using $\frac{\text{watts}}{\text{cost·date}}$ to determine outliers. Boxplot shows no outliers. Did this because it's basically cost efficiency relative to when it was made. When plotting each variable alone, 2 outliers for date, no outliers for cost and watts

b. mean = 461.5×10^5 = $462 \cdot 10^7$

sd = $170 \times 10^5 = 1.7 \times 10^7$

c mean(dataset\$cost/dataset\$MWatts) = $.57 \cdot \frac{100{,}000\text{ dollars}}{\text{MW}}$ = .057 \$/W

sd(dataset\$cost/dataset\$MWatts) = .19 → .019 \$/W

d The plot is slightly skewed left, with the frequency of plots dropping off sharply at .08\$/W. This is likely because people wouldn't be willing to spend any more than that per watt, so the generator wouldn't be built if it had a higher \$/W ratio. But technological constraints make it hard to decrease costs beyond a point, so frequency decreases as costs per watt approach zero.

2.12 Meat: mean calories: 158.7 sodium: 418.5

Poultry avg calories: 118.8 · avg sodium: 459

Beef avg calories: 156.9 sodium: 401.15

Poultry has the fewest calories, but highest sodium content

Meat is strictly worse than Beef, higher calories and sodium.

Beef has least sodium, and mediocre calorie content

I typed this page up because my handwriting was steadily deteriorating and this page was pretty messy

2.13

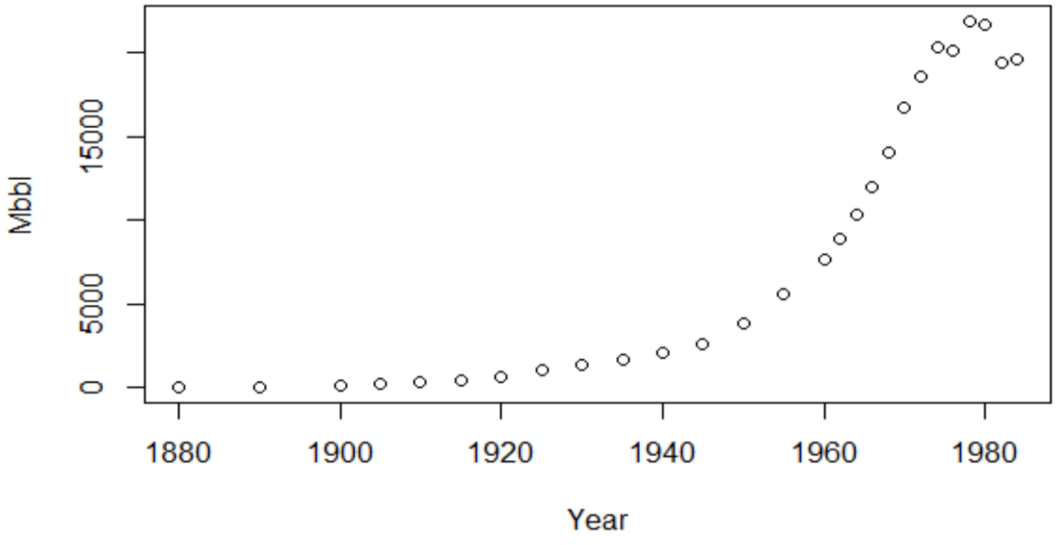
a. Group 3 has 4 outliers and the largest range. If outliers are eliminated, Group 3 has the lowest IQR range and variance. Middle 50% is bounded by 10 and 15 3+ syllable words. Groups 1 and 2 have no outliers, and a more even distribution. After eliminating the outliers of group 3, group 1 has the largest range, and group 2 has the second largest range. Group 1 has the highest median and mean of 3+ syllable words, Group 2 has the second highest in both median and mean, and Group 3 has the lowest median and mean of 3+ syllable words.
b. The three groups have very close median and means, but group 2 has the highest variance, closely followed by group 1. Group 3 has the lowest variance, but also has 3 outliers.

2.14

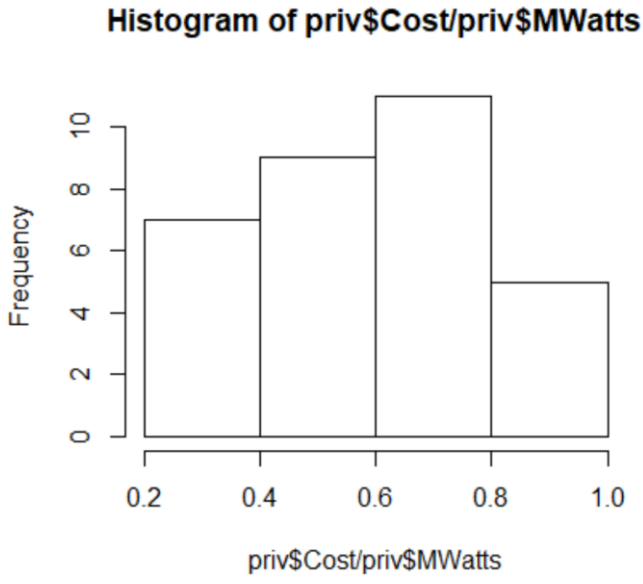
a. (note: higher ranked means a better school, but refers to lower numerical value in the rank column) Average debt trends downward for higher ranked schools, and the effect continues when you split the top third of private schools into thirds again. For the sake of interest, I also plotted the cost of attendance, and cost was positively correlated with rank. However, highly ranked schools actually gave much more need based aid, so the plot of cost minus aid actually trended downward for highly ranked schools, as shown in my plots.
b. Std(private cost) = 10461, std(public cost) = 4500. Private school tuition has a much higher value and range, so it's analogous to multiplying public tuition by a scalar, naturally increasing variance and standard deviation because the range is higher.
c. Each third shows that tuition is positively correlated with rank
d. Same as private schools, rank increases with tuition.

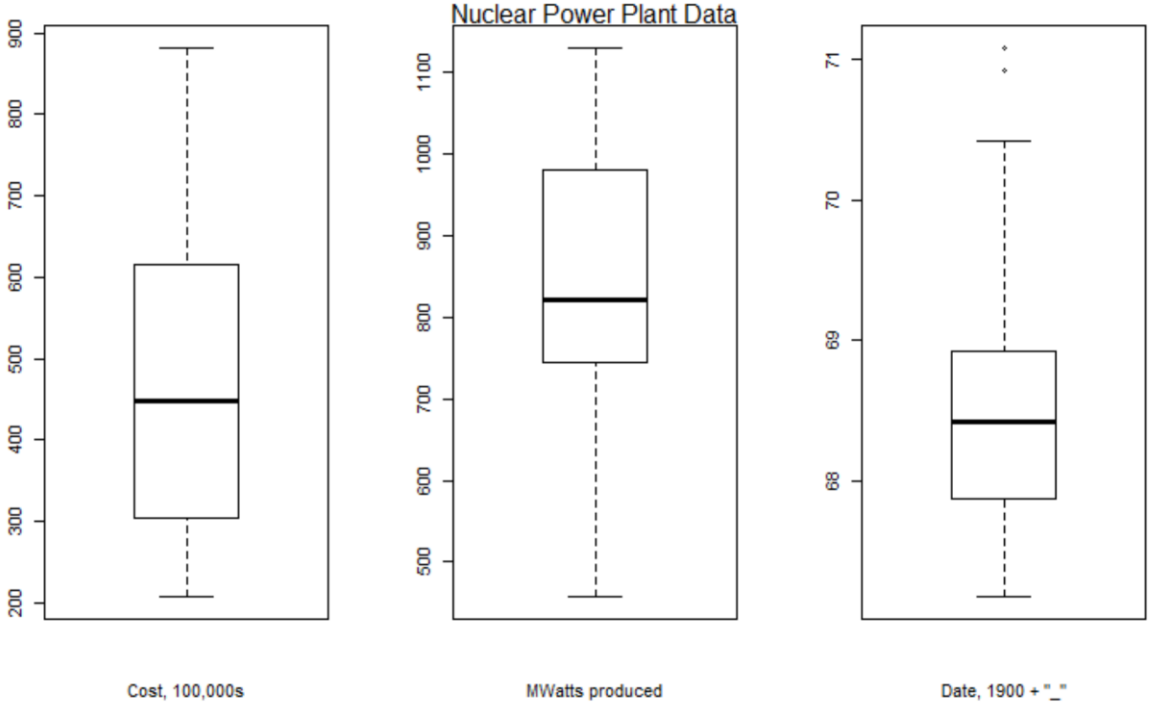
All the plots I referenced

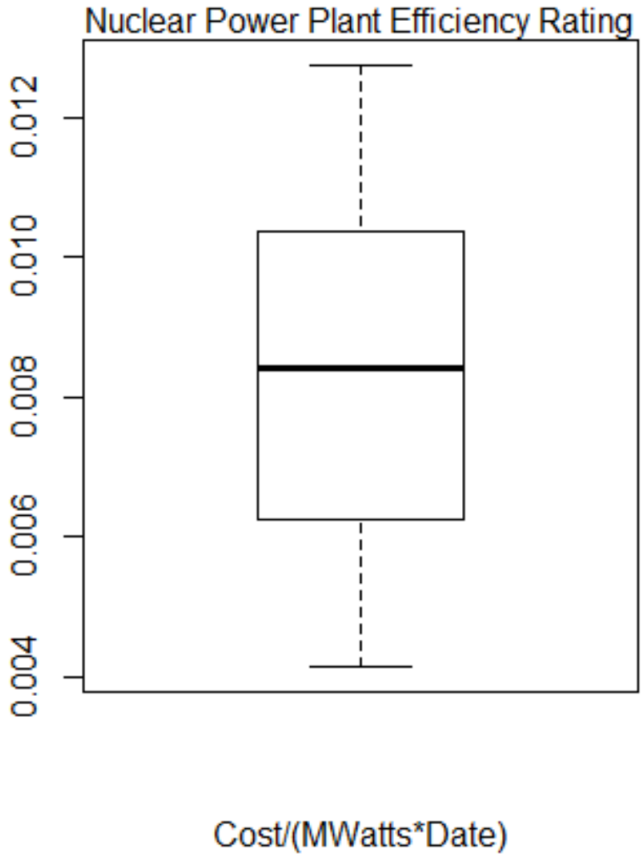
2.10



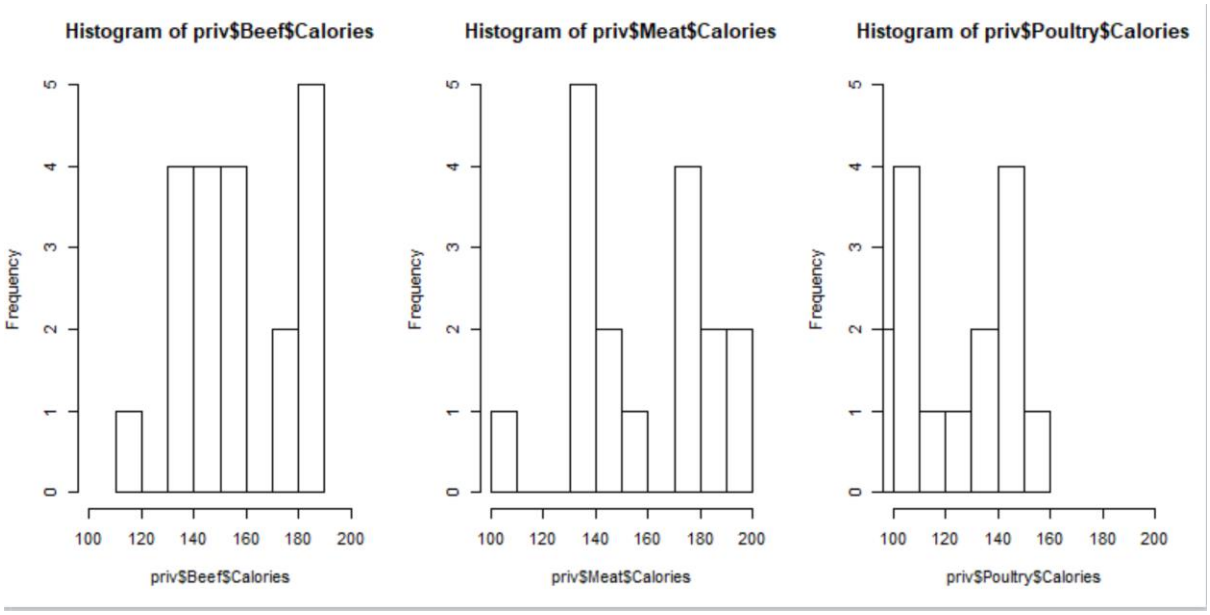
2.11

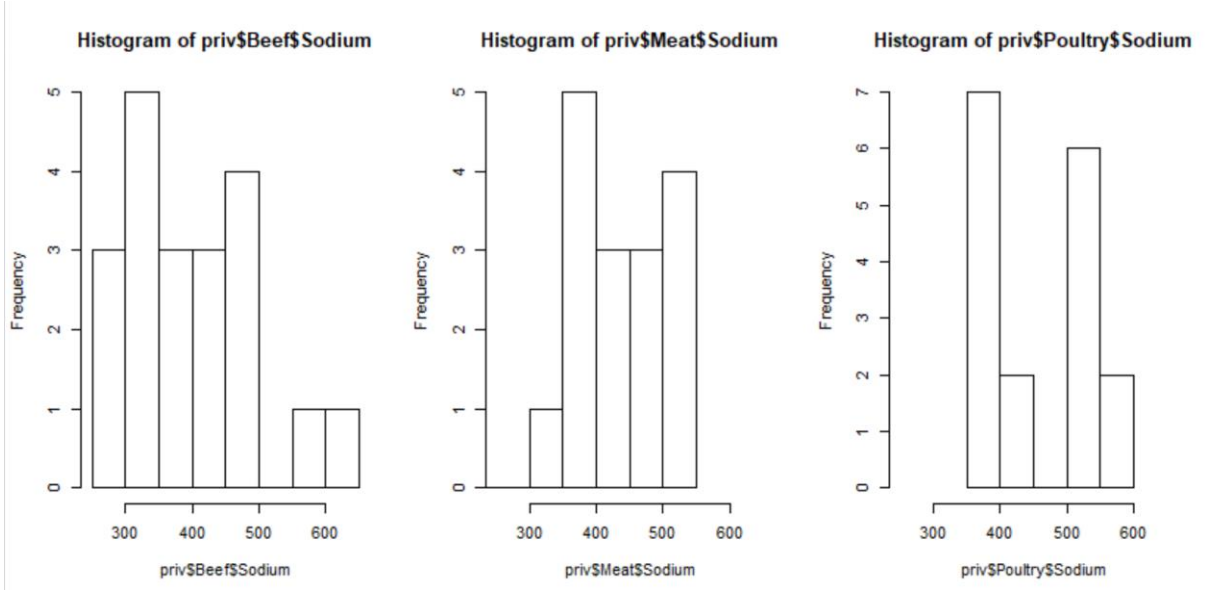

Histogram of priv$Cost/priv$MWatts
Frequency
0
2
4
6
8
10
0.2
0.4
0.6
0.8
1.0
priv$Cost/priv$MWatts


Nuclear Power Plant Data
200
300
400
500
600
700
800
900
500
600
700
800
900
1000
1100
68
69
70
71
Cost, 100,000s
MWatts produced
Date, 1900 + "_"


Nuclear Power Plant Efficiency Rating
0.004
0.006
0.008
0.010
0.012
Cost/(MWatts*Date)

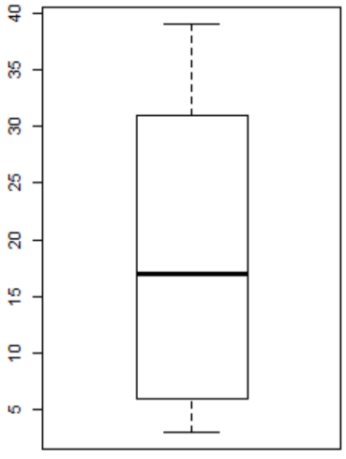
2.12


Histogram of priv$Beef$Calories
Frequency
priv$Beef$Calories
Histogram of priv$Meat$Calories
Frequency
priv$Meat$Calories
Histogram of priv$Poultry$Calories
Frequency
priv$Poultry$Calories

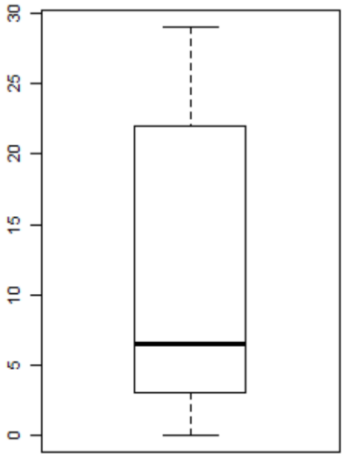

Histogram of priv$Beef$Sodium
Frequency
priv$Beef$Sodium
Histogram of priv$Meat$Sodium
Frequency
priv$Meat$Sodium
Histogram of priv$Poultry$Sodium
Frequency
priv$Poultry$Sodium

2.13

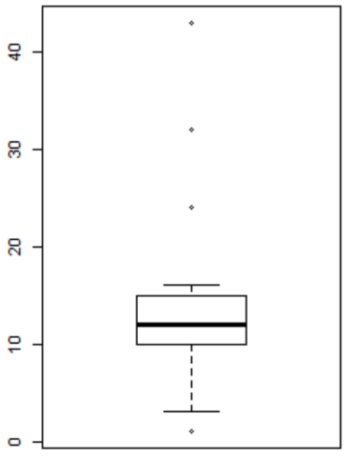
First graph is 3 + syllable words by group



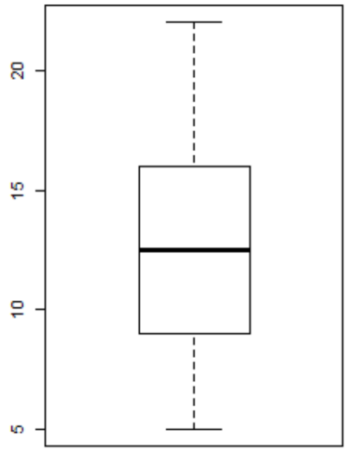
Group 1



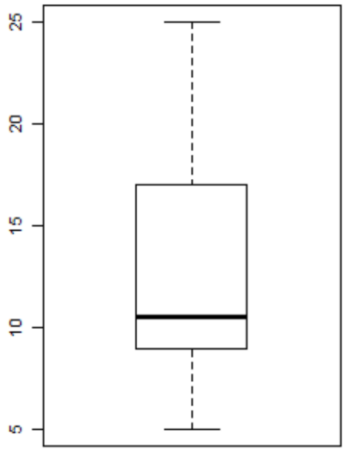
Group 2



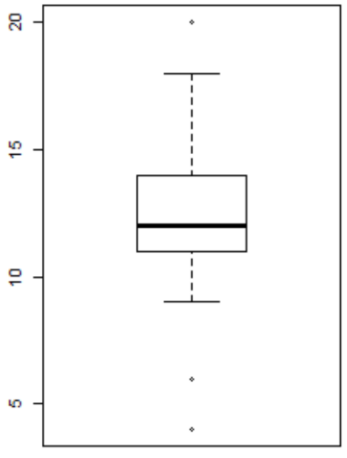
Group 3



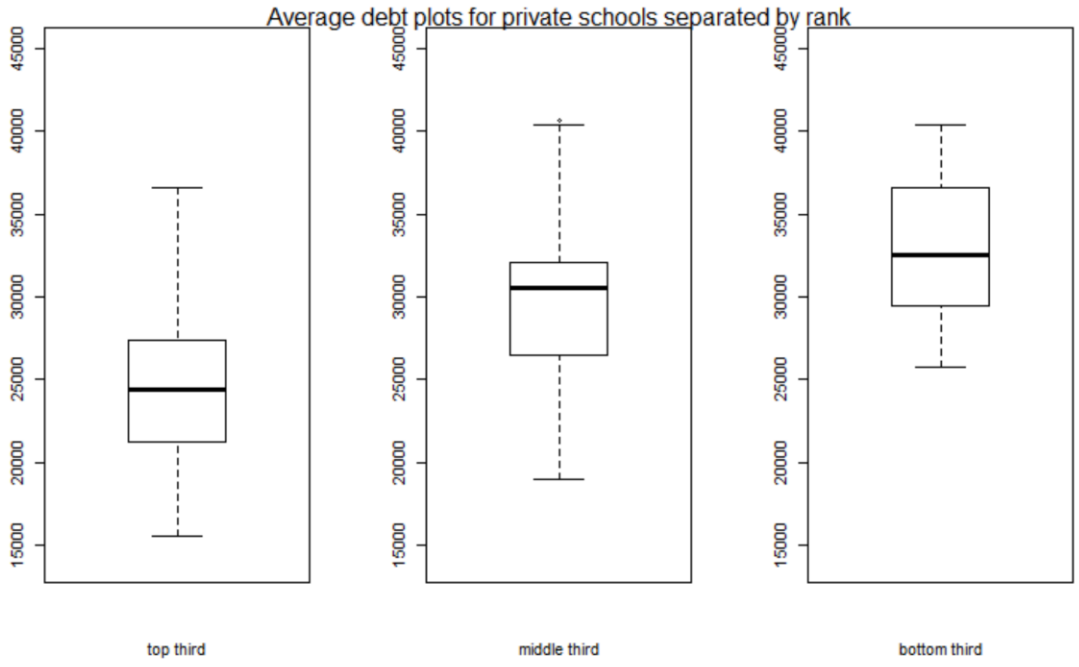
Group 1, ad sentences



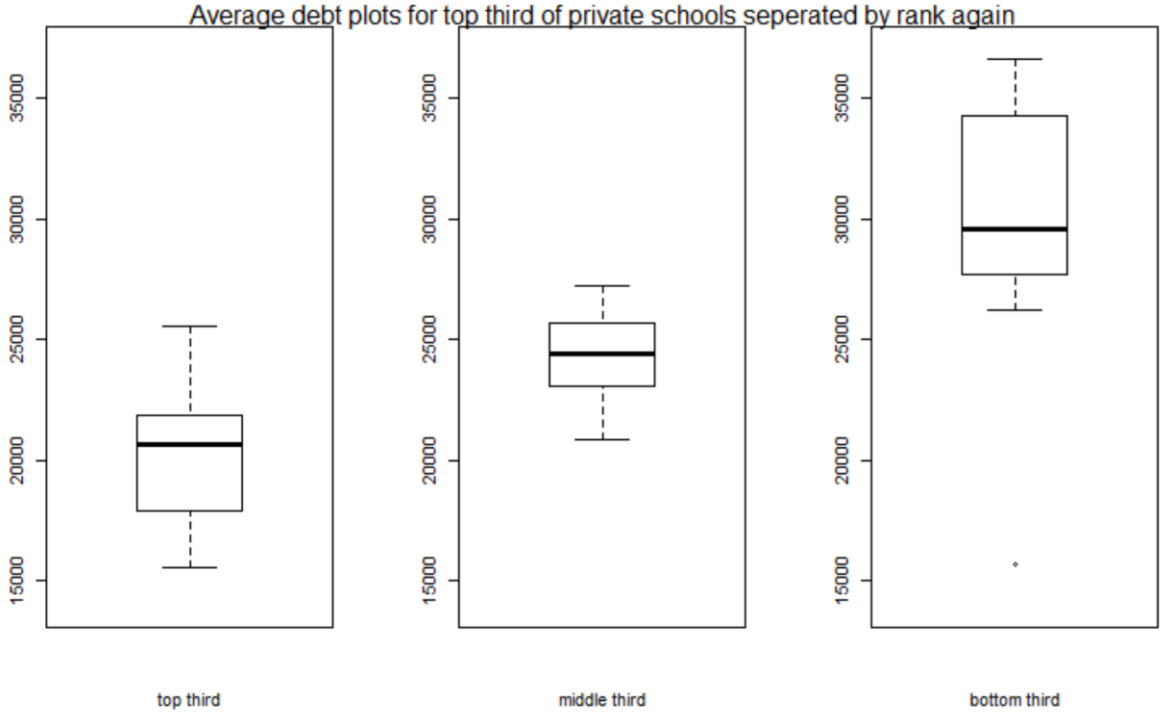
Group 2, ad sentences



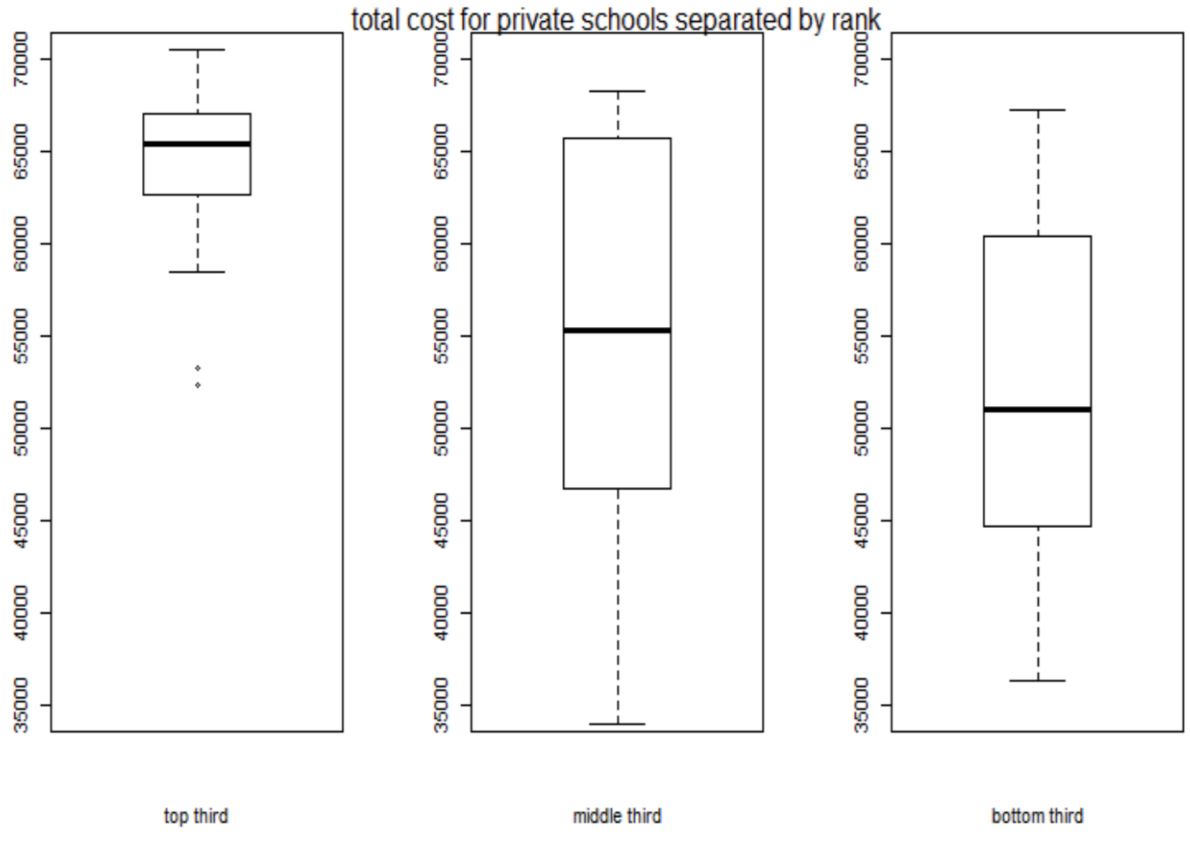
Group 3, ad sentences

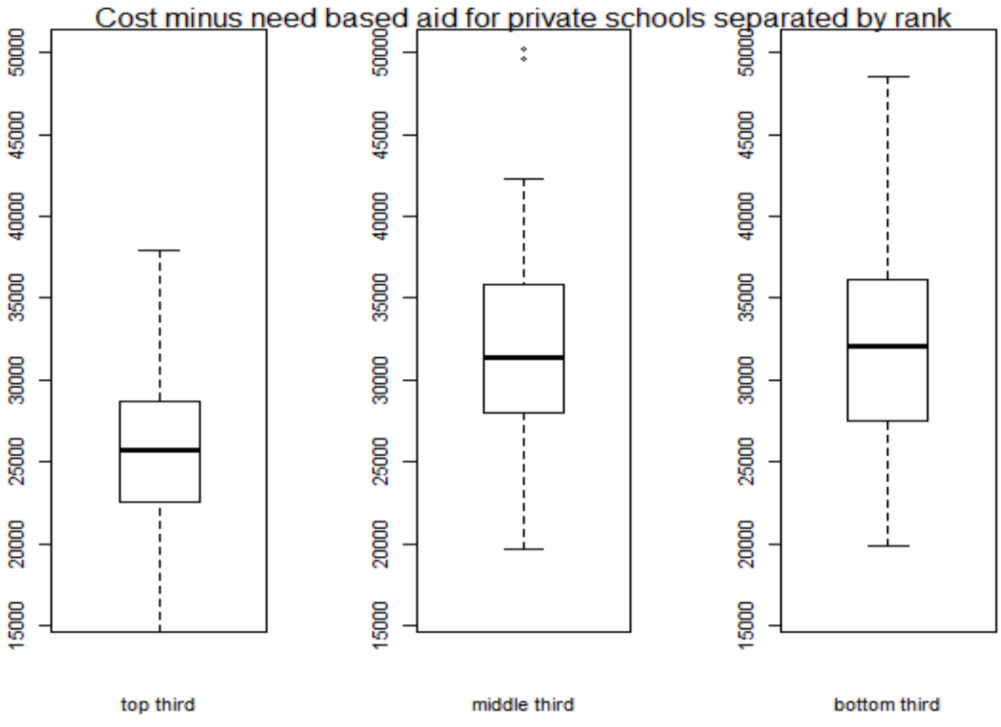

Average debt plots for private schools separated by rank
45000
40000
35000
30000
25000
20000
15000
top third
middle third
bottom third

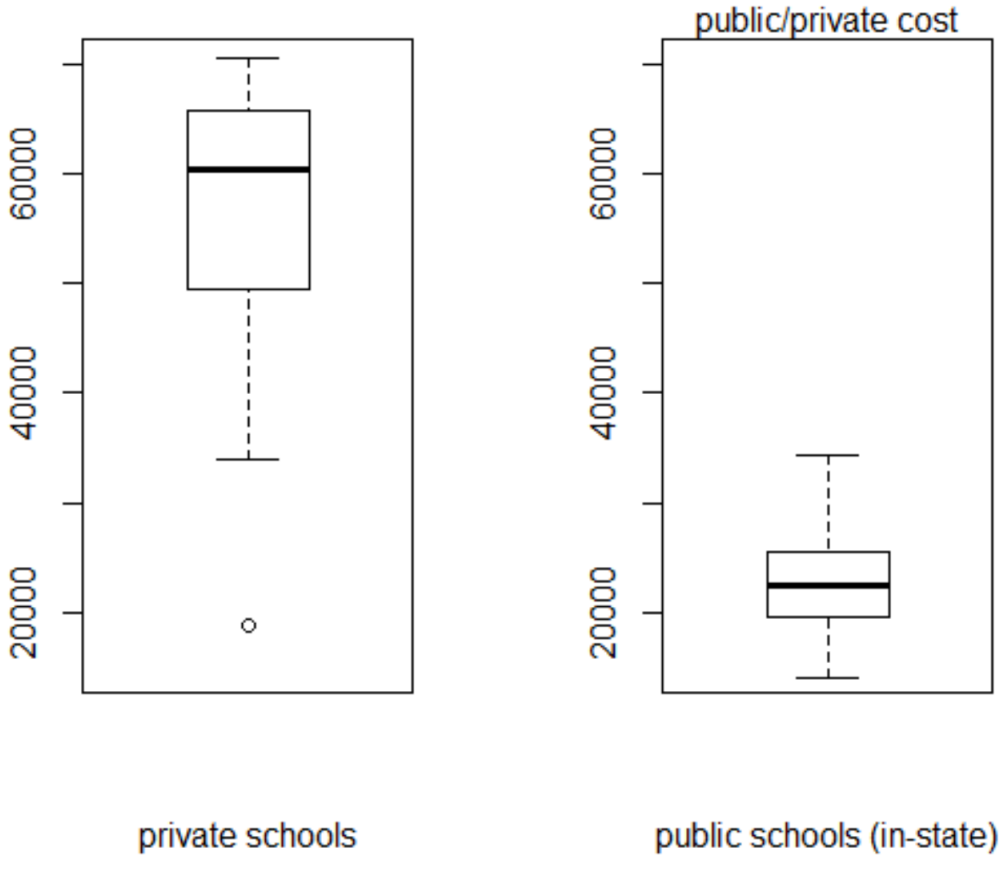
2.14

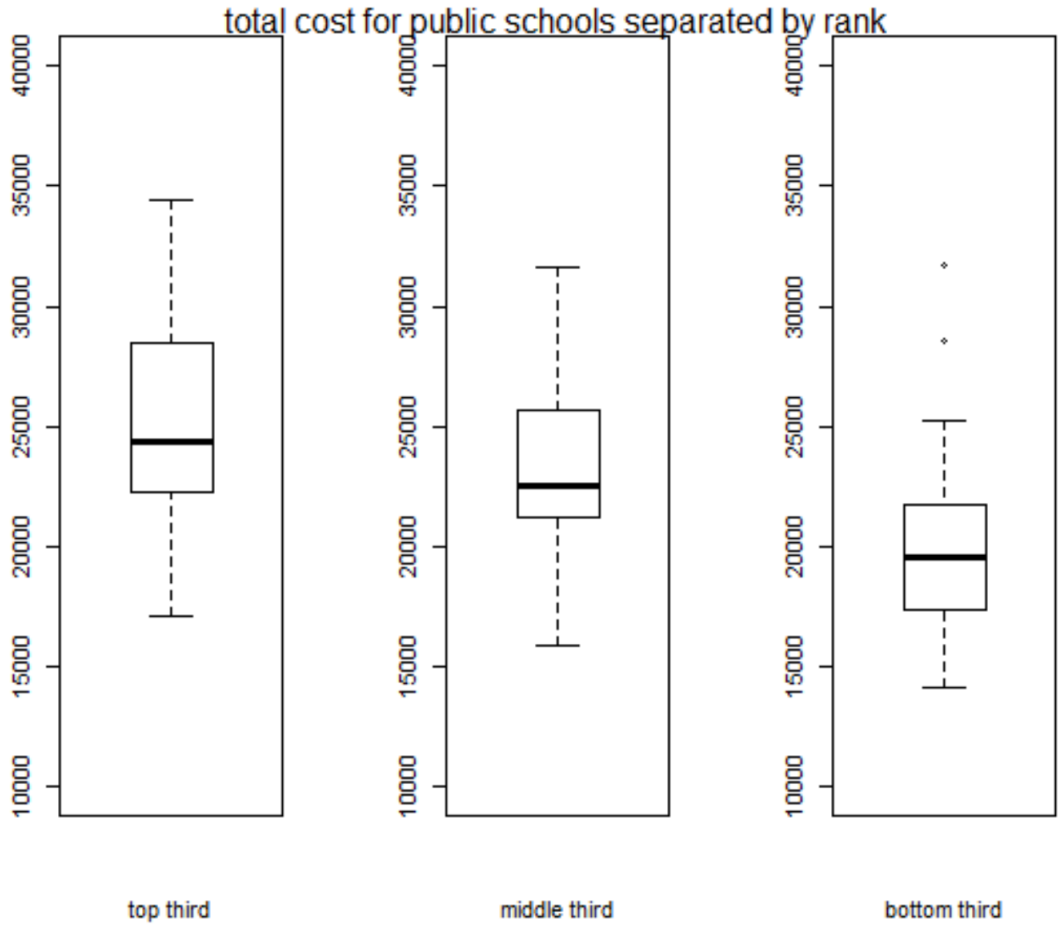

Average debt plots for top third of private schools seperated by rank again
35000
30000
25000
20000
15000
top third
middle third
bottom third

Just straight up cost




Cost minus need based aid for private schools separated by rank
15000
20000
25000
30000
35000
40000
45000
50000
top third
middle third
bottom third


public/private cost
60000
40000
20000
60000
40000
20000
private schools
public schools (in-state)



Similar trend as that observed for private schools, just less pronounced due to the smaller range and variance.