36-201: Statistical Reasoning

1. [16 points] *Matching.*

   a. Which distribution(s) are skewed left? . . . . . [A] B C D E F

   b. Which distribution(s) are clearly bimodal? . . . . . A B C D [E] F

   c. To which data might you apply logarithms to improve normality of the distribution? . . . . . A [B] C [D] E F

   d. For which unimodal distribution(s) does the mean=median=mode, approximately? . . . . . A B C D E [F]

2. [16 points] *Fill in the blank.*

   a. The eye-color of each person in this class . . . . . Qualitative

   b. The number of hours you spend per day listening to WRCT . . . . . Quantitative

   c. The weight of the letters in a mailman's bag at the start of his/her route . . . . . Quantitative

   d. Your zip code . . . . . Qualitative

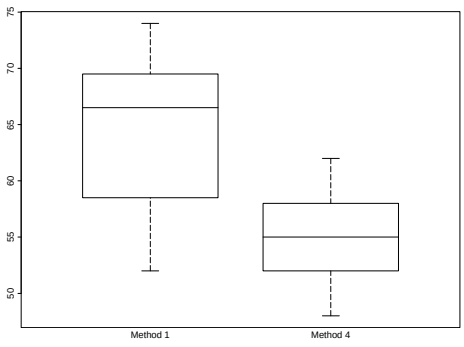
3. [22 points]

   a. [8 points] The sorted data and fences are:

```
Method 1:
            52 55 56 61 63 66 67 68 68 71 72 74
            Lower fence =  57.25 - 1.5*(70.25-57.25) = 37.75
            Upper fence =  70.25 + 1.5*(70.25-57.25) = 89.75

Method 4:
            48 50 52 54 55 55 55 58 58 62 62
            Lower fence = 52 - 1.5*(58-52) = 43
            Upper fence = 58 + 1.5*(58-52) = 67
```

   (A good way to sort the data is to make stem and leaf plots.) No data points are identified as outliers by the 1.5·IQR rule.



b. [9 points] Please circle the single best answer to each question, based on your boxplots:

(i) The distribution of car speeds under Method 1 is
unimodal. **bimodal.** **can't tell.**

(ii) There are outliers in the distribution of speeds for
Method 1. **Method 4.** **neither.**

(iii) The distributions of car speeds for these two methods differ mostly in
central location. **spread.** **both.** **neither.**

c. [5 points] The median speed is much lower for method 4 than for method 1—in fact, almost all the method 1 speeds are above the method 4 median.

The spread is also much less for method 4 than for method 1: for method 4 the IQR=6, less than half of the method 1 IQR, which is 13 (similarly for the range of the data).

Both data sets are approximately symmetric with no outliers, though some skewing to the right can be seen in the box for method 1.

Overall, method 4 appears to be more successful at reducing motorists speeds.

4. [25 points]

a. [5 points] The middle 50% of the PctVotingAge data lies between the quartiles, 50.60% and 60.60%.

b. [5 points]

- *Relations between the variables:* there is a mildly increasing relationship between the variables. It could be linear, but the data is so spread out it is a little hard to say.
- *Clustering:* There is no obvious clustering.
- *Outliers:* There might be outliers in the lower-left (Wash DC) or upper right (Minnesota). Washington DC is a little different from the states so it's not really surprising to see it as an outlier. I'm not sure what's going on with Minnesota.
- *Unequal Variability:* There is not much evidence of this in the graph.

c. [5 points] Wisconsin's PctVotingAge = 66.0. So the estimate of PctRegistered would be

```
PctRegistered = 61.6 + 0.228*66.0 = 76.6, approximately
```

d. [5 points] $r = 0.27$, so $r^2 = 0.0729$, so about 7.3% of the variability in PctRegistered is explained by a linear function of PctVoting Age.

e. [5 points] The r^2 is quite low, so a straight line fit doesn't capture much of the variability in PctRegistered.

On the other hand, there isn't much evidence of a nonlinear fit in this data (students who try median trace plots will see a little bit of curving but not much).

The main culprit seems to be the huge variability in PctRegistered across the states, which dilutes the relationship between the two variables.

5. [21 points] Table 1 below shows the educational level of US adults of various ages, in 1994.

	25–34	35–64	65 and up	Total
Did Not Complete HS	5705	14152	11561	31418
High School	14472	31539	10504	56515
College, 1–3 Yrs	11913	19107	4853	35873
College, 4 Yrs or More	9816	22887	3843	36546
Total	41906	87685	30761	160352

Table 1: Number (in 1000's) of US adults completing various educational levels, cross-classified by years of age.

a. [4 points] Which is the explanatory variable and which is the response? (mark one box only)

☐ Educational level is explanatory and Age is response.

☒ Age is explanatory and Educational level is response.

b. [6 points] Tables 2 and 3 contain the row percents and column percents for this data, but some cells are missing. Fill in the missing cells.

	25–34	35–64	65 and up	Total
Did Not Complete HS	18.16	45.04	36.80	100.00
High School	25.61	55.81	18.59	100.00
College, 1–3 Yrs	33.21	53.26	13.53	100.00
College, 4 Yrs or More	26.86	62.63	10.52	100.00
Total	26.13	54.68	19.18	100.00

Table 2: Row Percents.

	25–34	35–64	65 and up	Total
Did Not Complete HS	13.61	16.14	37.58	19.59
High School	34.53	35.97	34.15	35.24
College, 1–3 Yrs	28.43	21.79	15.78	22.37
College, 4 Yrs or More	23.42	26.10	12.49	22.79
Total	100.00	100.00	100.00	100.00

Table 3: Column Percents.

c. [6 points] In Tables 4 and 5, one expected count is missing and one standardized residual is missing. Fill in the missing expected count and standardized residual. Show your calculations here.

	25–34	35–64	65 and up
Did Not Complete HS	8210.70	17180.25	6027.05
High School	14769.49	30904.00	10841.51
College, 1–3 Yrs	9374.96	19616.37	6881.67
College, 4 Yrs or More	9550.84	19984.38	7010.77

Table 4: Expected Counts.

	25–34	35–64	65 and up
Did Not Complete HS	-27.65	-23.10	71.28
High School	-2.45	3.61	-3.24
College, 1–3 Yrs	26.21	-3.64	-24.45
College, 4 Yrs or More	2.71	20.53	-37.83

Table 5: Standardized Residuals.

d. [5 points] Every number in the table of standardized residuals is huge compared to our rules of thumb (± 1.5, ± 2), and some are enormous (-27 or -37, $+71$, etc.). So there is very strong evidence against age and educational level being independent.

Another interesting thing to notice is that for the "Did not complete HS" row, the standardized residuals go $-$, $-$, $+$, and for the "4 or more years of college" row they go $+$, $+$, $-$. This says that too many older people did not finish high school, relative to independence, and too many younger and middle-aged people did complete four or more years of college. You can see this in the table of column percents as well (which makes sense to look at, since Age is explanatory).

The overall pattern, that the older you are the less school you are likely to have finished, makes sense, because higher education was less available to people who are older today, when they were younger and could have taken advantage of it.