1. [3 points] IPS: 2.18
   (a) Scatterplot:

   ![Bone Strength: Scatterplot of Dominant ~ Nondominant](image)

   Bone Strength: Scatterplot of Dominant ~ Nondominant

   Nondominant Arm
   Dominant Arm
(b) & (c) & (d) & (e): The pattern is a straightforward positive linear relationship, with high (positive) correlation, and we see that in general, dominant arm strength is slightly above nondominant arm strength. There are no striking deviations from this pattern, and no outliers (the high point on the top right is not an outlier to this pattern).

Two points for the scatterplot: one for the data, other for correct axes / labelling of axes. One point for a generally good description from (b) through (e): doesn’t have to match the above exactly, but can’t be too far off.

2. [2 points] IPS: 2.43
   
   (a) We have a correlation of 0.905.
   
   (b) Yes: the relationship is reasonably linear with no real outliers, and is homoskedastic. Correlation here is a very good numerical summary of this plot.

   One point for each part.

3. [3 points] IPS: 2.67
   
   (a) We won’t replot this: it’s the same plot as Q1 (a).
   
   (b) Scatterplot with regression line:
(c) Our answer will be similar to the description in Q1: Nondominant strength is a strong linear predictor of dominant strength, with homoskedasticity and no outliers here. Dominant strength is generally slightly better than nondominant strength.

Two points for scatterplot with regression line (the text of the regression, i.e. “\( y \sim 2.739 + 0.936x \)”, is not required), one point for a reasonable description based on the plot (doesn’t have to match the above).

4. [4 points] IPS: 2.160

(a) Here, we’ve included a line between the points, you don’t need this. Also note the reversed Y axis: this is because we’re plotting
ranks, so lower is better - you did not have to reverse the plot either.

(b) The regression is \( \text{Rank} \sim 119,559.7 - 59.2 \times \text{Year} \). Now with regression line (again, you don’t need a reversed axis, and you don’t have to have printed the numeric regression equation):
(c) More popular! The rank is decreasing, at a rate of nearly 60 ranks per year (so indeed we can expect Atticus to be the most popular name soon after 2019...).

One point for the first scatter plot, one point for getting the correct regression equation, one point for adding it correctly to the plot (if the equation is wrong, but the plot line matches the wrong equation, the point for the plot line can still be earned), and a final point for a reasonable answer to (c).

5. [3 points] IPS: 2.170

(a) First, let’s replot it, as the description indicates the data has
changed slightly, and we can add in visual indications for Utah and California (and a regression line for good measure):

![Graph showing the relationship between smoking and fruit/vegetable consumption.]

There is a moderate negative association between the two - we’d expect this from the perspective of healthier people / states eating lots of fruit and veg and avoiding cigarettes. Also, smoking correlates positively with poverty, while healthy eating correlates negatively with it.

(b) The correlation is $-0.491$ - moderate (and of course negative...).

(c) Utah is marked on the above plot. Utah has the lowest smoking percentage of any state (probably religious reasons), but an average consumption of fruit and veg - lower than we’d expect, based
on its smoking percentage. We might expect this because religion probably affects Utah’s smoking percentage a lot, but has limited effect on its fruit/veg consumption.

(d) California is marked on the map above. California has the second lowest smoking percentage, barely above Utah, with both quite a bit below the next lowest (Hawaii, 12.1%). It has a high percentage of fruit/veg consumption, fifth highest. This high consumption is about in line with what we might expect based on its low smoking percentage.

(e) Answers will vary for this. Some interesting cases on the unhealthy side are Kentucky (probably the least healthy: clearly most smokers, and nearly the least in fruit/veg) and Oklahoma (least fruit/veg, a lot of smokers too); in contrast Vermont is doing very well with fruit/veg, and is also the state with the largest residual when fruit/veg is regressed on smoking (“worst predicted”).

One point for the correct correlation. One point for a reasonable description of both Utah and California’s relative position. One point for either a good answer in (a) (note: plot not required) or (e).

6. [3 points] The intercept is $-167$. It’s the expected value of $Y$, here weight, when $X = 0$, which in this case means inches $= 0$. Here this really has no meaning since we never have $X = 0$ in our data. The slope is 4.7, which here means that the expected value of $Y$, weight, increases by 4.7 pounds for every extra inch in $X$, height.

One point each for correctly identifying intercept and slope. One more point for both interpretations being correct.

newborns.

Two points for correct predicted value. One point for reasonable interpretation.

8. [3 points] Our three scatter plots:
The strongest relationship is with weight, since that has the tightest relationship with miles per gallon (least spread). Actually that’s not too obvious, so you might want to calculate correlations (or similarly, $R^2$ values) to check...

One point for each correct plot. Remove a point if justification for strongest relationship isn’t given (actually it’s OK to choose the wrong graph, but the “why” is important).

9. [3 points]
   
   a Yes, they’re positively correlated.
   
   b Yes, they’re positively correlated.
c No, causation isn’t supported by the observational data given.

d No: again no causal relationship. [However this statement might remind us of a somewhat similar statement: 11.56% ($R^2$ is correlation squared) of the variance of performance is non-causally explainable by study time]

e Not really; 0.34 isn’t too big

Takes points away from 3 for every wrong or badly supported answer; note that some may justify “yes” for e well - if good justification is given, that’s fine.

10. [3 points] The boxplots:
For SAT, we see that the male students score a lot higher, with about the same spread.

For GPA, the two seem to be about the same, with minor differences depending on what your favourite numerical summary might be.

One point for each correct plot, and lose a point if gender not labelled. One point for a reasonable set of interpretations.