Glenn Clapp
For each block of R code state what's being done and why.
i. what does the code do
ii.why it is (or is not) needed
iii. interpret any output.

Question 1:
a. This block reads the data in from a *.csv and factors it - changes each level # into a string describing the size of a stroke lesion. Finally, the re-factored table is printed to stdout as shown in the output for a. Not much to interpret here, mostly house-keeping preparing for analysis.

b.
i. This section is building a linear model with a dependent variable "Scores" and independent variables "Side" and "Size" The asterisk tells the model to include the reaction term seen in the output "Side:Size"
ii. This is needed because the linear model is what will be analyzed in the coming section, and the reaction term is needed because one of the underlying assumptions of the ANOVA is that there is no significant interaction term. The anova on m1 does not reveal a significant difference between any of the groups.
iii. $p = 0.6408$, for the model, no significant interaction terms and insufficient evidence to suggest that any of the means differ.

c.
i. In this block of code a levene test is being run on the data, first with the mean as a measure of center and then with the median as a measure of center.
ii. The levene's test is to check for equal variances, which is another underlying assumption of the ANOVA. It is however not necessary to run both of these. The mean is appropriate for normally distributed data, and the median is appropriate for non-normally distributed data.
iii. There was no test for normality in this example. However, the null hypothesis is that the data are normally distributed, and the sample sizes are large enough that we'd be unlikely to reject. So the mean should be used. Because of this, we reject the null hypothesis that the variances are equal.

d.
i. This section is checking residuals.
ii. The assumption that there are equal variances implies that the residuals will be $\sim N(0,\sigma^2)$ where the variance is equal to that of the populations.
iii. This result should agree with the levene's test. This code outputs a graphic though which makes it obvious why we failed the levene's test which is an extreme outlier for which the residual is over 150. It might be cause to go back and use the median in the levene's test in which

case we would fail to reject the null of equal variances. Additionally, the intercept is being forced through the origin which should not be done in this case.

e.
i. this code builds a new data set (m5) eliminating data points over 100. and performes an anova. Next it builds m3 using the new data set but designating SIZE as a blocking factor.
ii. one model leaves in the interaction term and the other considers size a blocking factor. m3 is the useful model here as it will allow us to determine if there is a difference in awareness given the side of the head and ignores the effect we know is there of size on awareness.
iii. The output here shows that there is a significant effect based on SIZE, but not SIDE in either model. The residuals do seem to have improved however.

f.
i. this code is running an anova after defining m4 which does not consider SIDE with the modified data set that doesn't have the extreme outlier.
ii. This test is interesting as we've determined that SIDE does not seem to be a contributing factor even when SIZE is considered a blocking factor, so it's removed from the model.
iii. The ANOVA shows a significant effect due to SIZE on awareness.

g.
i. This is re-running the levene's test on m4 which does not consider SIDE and tests residuals
ii. It is necessary to test underlying assumptions
iii. There appears to be a "fanning" pattern, so we should follow up with a non-parametric test.

h.
i. Non-parametric test
ii. Necessary to deal with non-normally distributed data or residuals.
iii. The Kruskal-wallis test did not find a significant effect.

i.
i. Post-hoc testing with a bonferroni correction to reduce the chance of type I errors. This is necessary as repeated t-tests can result in relatively high probabilities for type I errors.
ii. This would be necessary if the analysis of variance tests had revealed a difference in groups as the ANOVA or Kruskal-wallis do not give information on which groups were different, just that not all means were equal. However, the Kruskal-wallis test did not find this, so post-hoc tests are not necessary
iii. As expected, no significant differences between groups were found. Thought the difference between "5cm+" and "1-3cm" was almost significant with a p-value of 0.099

Question 2

a.
i. Reading in data and plotting a scatterplot
ii. CVDIG1 is the dependant variable and AGE and ARMLENGTH are independent
iii. Self-explanatory, this is useful for visualizing relationships.

b.
i. Non-parametric regression tests. Several different methods are used
ii. It is only necessary to pick one of these methods.
iii. All models seem to show a strong, positive correlation between CVDIG and Arm length and a weak, positive correlation between CVDIG and AGE. Perhaps obviously, no correlation between age and arm length as all test subjects were adults.

c.
i. 4 linear models are created and a dataframe is made containing data on the r^2 and adjusted r^2 values for each of the models.
ii. This is a pretty efficient way to analyze a lot of models.
iii. It appears that the best results were obtained with m2 if we look at the adjusted r^2. I think this is appropriate as the adjustment is for over-fitting and m2 includes Age as an explanatory variable and armlength as a blocking factor. This represents the most likely model of the situation intuitively, and the data seem to back that up.

d.
i.Summarize m2
ii. Necessary to get the actual equation
iii. Repeat of some of the same information as part c with the addition of coefficients and an intercept so that we can determine the actual model.

CV = 0.035AGE + 0.197ARMLENGTH -1.946

e.
i.Testing residuals
ii.Necessary to test assumptions
iii.Good fit.

f.
i.Levene's test. No result for the commented out line because there is only one replicate for some armlength, age pairs?
ii. This test is not necessary as it tests for equal variances which is not an underlying assumption in this case.
iii. Unnecessary test.

g.
i. This code builds confidence and predition intervals based on m2 at the specific point arm length of 76 and an age of 63.
ii. Interesting if that data pair means something to you. Perhaps a researcher's arm length and age?
iii. The output is a PI, an interval for observations at an arm length of 76 and an age of 63, and a CI, an interval for the regression line developed by another sample of similar size being drawn from the same population.

h.
i. Here we have a new model that only considers arm length.
ii. This is the strongest predictor of CV based on our earlier analysis, so it might be interesting to ignore age.
iii. This model is not as good as m2. The unexplained variation is higher though the p-value suggests that the model is highly significant. Residuals are also tested and it looks as though the relationship between armlength and CV may not be linear.

i.
i. prediction and confidence intervals are created and plotted.
ii. Good visual for showing these intervals.
iii. The small interval around the regression line is the CI and the large one that seems to bound the observations is the PI. There do appear to be a few data points outside the PI.

j.
i. A point, horizontal, and vertical line are added to the plot.
ii. Could be necessary depending on what the data is being used for. Perhaps illustrating a specific observation in this case.
iii. An armlength of 103 corresponds to a CV of 19 in a particular observation which puts it below the regression line, outside the CI but within the PI. So, nothing abnormal there.

k.
i. 45 degree labels are added to the data points that fell outside the PI
ii. This could be useful to identify points in a large data set that fall outside the PI
iii.

l.
i. predict a CV confidence interval for an armlength of 50.
ii. good for point predictions.

iii. a point estimate of 9.23 with a lower and upper bound of 8.472 and 9.992 respectively. This estimate is for the point through which the regression line passes.

The second estimate is a PI and it yields the same point estimate with a lower and upper bound of 7.574 and 10.891 respectively for future observations at an armlength of 50.