Question one:

True or False:

a) In linear regression if we increase the training set size, the mean training error will decrease.

False. As we increase training set size the training error increases since number of points increases therefor our error also increases. But our generalization gets better.

b) In linear regression, the mean of residuals is always zero.

True. Summation of residuals is zero therefore the mean is also zero.

c) Standardization of features is required before training a logistic regression model.

False, Is not required. If logistic regression with LASSO or ridge regression is used standardization is required.

d) Correlated variables can have zero correlation coefficient.

False. Correlated variable has non zero correlation coefficient.

e) In Ridge regression if you apply a very large penalty, some of the coefficients will become absolute zero.

True. Large penalty make some variables absoulty zero.

f) If R2 = 0 then there is no relation between x,y.

False. There is no linear relation. For instance, consider $y = x^2$. R in this case will be zero but there is a relation.

g) If = 0.791, there is a strong positive linear relationship R with strength 2 0.88 between variables.

False, the sign matters it would be false, since $R = \pm 0.889$

h) The residuals measure the distance between the observed value and the regression line.

False! residual is vertical distance between the observed value and the regression line.

Question 2

A.

$$b_1 = \frac{3.9983 - 4.010}{-0.0833} = 0.14, \qquad T = \frac{0.14}{0.032} = 4.34$$

B.

We use hypothesis testing in order to answer this question.

Null hypothesis: $\beta_1 = 0$

Alternative Hypothesis: $\beta_1 \neq 0$

$$T = 4.34, df = 461$$
$$P_{value} \approx 0$$

Null hypothesis is rejected therefore data provide convincing evidence that is strong relationship between the evaluation score and adornment

C.

$$t_{461} = -1.96$$
$$Confidence\ Interval: 0.14 \pm 0.063$$

Yes, H_0 is rejected and the confidence interval does not include 0.

Question 3

A.

$$Blood\ pressure = -80.41 + 0.44 * smoke - 3.33 * excercise\_hour - 0.01 * age + 1.15 * height + 0.05 * weight - 8.40 * water\_consumption$$

B.

Smoke: The estimated blood pressure of smoker people is 0.44 unit higher than non smoker people.

Water consumption: The model predicts a 8.44 unit decrease in the blood pressure for people who does not smoke.

(Interpterion above is only valid if other factors hold constant)

C.

$$R^2 = \frac{83.29}{332.57} = 0.2504$$

$$R_{adj}^2 = 1 - \left(\left(\frac{249.28}{332.57}\right) \times \left(\frac{1235}{1229}\right)\right) = 0.2448$$

Question 4

a. Null Hypothesis: chance of admission of three groups (good, bad, medium) is same.
Alternative Hypothesis: at least one pair of group has different chance of admission.

b.

	Df	*Sum Sq*	*Mean Sq*	*F Value*	*Pr(>f)*
Class	2	1144	572	6.1516	0.002326
Residual	425	39518	92.98		

$$Sum\ sq\ class = 572 \times 2 = 1144$$
$$Mean\ sq\ residual = \frac{39518}{424} = 92.98$$
$$F\ Value = \frac{MSG}{MSE} = \frac{527}{93.2} = 6.1516$$

Since Probability(>f) = 0.002326<0.01 We can reject null hypothesis.

c. We t test in order to find out.
T stat computes from following:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{SE}$$
$$SE = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{92.98^2}{200} + \frac{92.98^2}{150}} = 10.04$$
$$t = \frac{8}{10.04} = 0.8$$
$$p\ value = 0.32$$

Doesn’t give enough evidence to reject null hypothesis.

Question 7

A.

Mean = 34

Sd = 14.63

Median = 30.5

B. In Bootstrapping Ali should sample with replacement from original sample. He can increase number of his data point and obtain a new dataset with this method. He can use this method for inferencing on estimators that CLT does not apply such as median.

Bootstrap does not have strict rule on sample size, and our data has not extreme values. On the other hand we know distribution of holes in sangak bread has finite variance and bootstrapping can be useful.

C. Shape of bootstrap density should be approximately symmetric, center should be some where near statistic calculated from main sampled distribution.

D. Median is my parameter of interest. Estimation -> 30.5

E. $4.85 \times 1.96 = 9.506 \rightarrow 95\ \%\ confidence\ interval :$ (21, 40) or 30.5 ± 9.5

Question 8

A. P value correction let us perform multiple tests on same data. If multiple tests perform on a same dataset chance of being randomly significant (less than 5%) increases, therefor correction is needed. To be clear consider this example: Researchers wants to find out whether eating dark chocolate affect blood pressure, blood oxygen level, body fat and some other factors or not. They can not perform multiple tests for each factor on same group of people since one factor might be randomly significant because of multiple comparison.

B. Bonferroni Correction: After performing p value it should be compared with significance level which has been divided by number of tests. It is more conservative and focuses on keep type 1 error down.
Benjamin- Hochberg Correction:
Put the individual p-values in ascending order.
Assign ranks to the p-values. For example, the smallest has a rank of 1, the second smallest has a rank of 2.
Calculate each individual p-value's Benjamin-Hochberg critical value, using the formula (i/m)Q, where:
i = the individual p-value's rank,
m = total number of tests,
Q = the false discovery rate (a percentage, chosen by you).
Compare your original p-values to the critical B-H from Step 3; find the largest p value that is smaller than the critical value.[1]
Benjamin Hochberg focuses on false discovery rate and is less conservative than Bonferroni. For large number of test Benjamin Hochberg is better since Bonferroni treat all p values equally but in small number of tests their performance is somehow equal.

C. The Bonferroni correction and Benjamin-Hochberg procedure assume that the individual tests are independent of each other. Also, the Bonferroni assumes that the p-value provided follows a uniform distribution under the null hypothesis[2].

D. If correlation coefficients, there is no need to do any correction.

[1] https://www.statisticshowto.com/benjamini-hochberg-procedure/
[2] https://stats.stackexchange.com/questions/156584/bonferroni-adjustment-and-assumptions#:~:text=The%20Bonferroni%20(and%20similar%20corrections,your%20experiment%20again%20and%20again.

E. 1/10 * 0.05 = 0.005

P – value	Rank	Q
0.0008	1	0.005
0.009	2	0.010
0.165	3	0.015
0.205	4	0.020
0.396	5	0.025
0.45	6	0.030
0.641	7	0.035
0.781	8	0.040
0.9	9	0.045
0.993	10	0.050

Only two first column is acceptable due to Benjamin correction.