Part 1

1-- Fully interpret the coefficient on Education in Model 1 (remember that “fully interpret” means that you consider the metrics that X and Y are measured in; you note other variables in the model; you note statistical significance; you make clear that you are predicting “on average”; and if it is an indicator variable, you note the reference or omitted category used) (12 points).

Interpretation: educ coeff = 0.152
For every 1 unit increase in education (corresponding to an additional year of schooling), on average, someone's liking of classical music increases by 0.152 units. This means that individuals with higher levels of education tend to have a greater liking for classical music, all else being equal.

Metric:

- Independent Variable: Education (measured in years)
    - Education is measured in years of schooling. Each unit increase in education corresponds to an additional year of schooling.

Dependent Variable: The dependent variable, Liking of Classical Music, is measured on a scale of 1 to 5, ranging from "very much do not like it" (1) to "very much like it" (5).

Other Variables in the Model: The model includes another independent variable, Older, a dummy variable indicating whether the respondent is 50 or older. The intercept term is also included in the model.

On Average: The coefficient represents the average effect of a 1 unit increase in education on how much someone likes classical music, holding other variables constant.

Statistical Significance: The coefficient is statistically significant, as indicated by the t-statistic of 15.331 and the p-value of <0.000001. This suggests that the relationship between education and liking classical music is unlikely to have occurred by chance.

Indicator Variable: Education is not an indicator variable; it is a continuous variable representing the number of years of schooling.

2– As we can see, the researcher finds that the coefficient on Education in Model 1 is highly statistically significant. In general, what does it mean for a coefficient to be “statistically significant”? Why would the researcher say that this coefficient is “highly” statistically significant? Specifically, please make sure to explain the idea of statistical significance by talking about the concepts of standard errors, t-statistics, p-values and sampling distribution. (26 points)

When a coefficient is “statistical significance”

- It has a p-value of less than 0.05
- The estimated relationship between the independent variable and the dependent variable is unlikely to have occurred by chance
- Statistical significance is determined by evaluating several key concepts: standard errors, t-statistics, p-values, and the sampling distribution.

Standard error = 0.01

- Standard error = sampling distribution standard deviation
    - Approximates the population’s standard deviation, divided by the square root of the sample size.
- In regression analysis, standard errors measure the variability or uncertainty associated with estimating the coefficient. They indicate the average amount by which the estimated coefficient may differ from the true population value
- Smaller standard errors imply greater precision in estimating the coefficient, as they suggest that the estimated coefficient is close to the true population value
- The standard error associated with the coefficient on Education (0.01) represents the average amount by which the estimated effect of education on the liking of classical music may differ from the true population effect.

T-Statistics = 15.331

- The t-statistic (15.331) is calculated by dividing the estimated coefficient (0.152) by its standard error (0.01).
- It measures the magnitude of the estimated effect relative to its uncertainty.
- In this case, the large t-statistic indicates a strong relationship between education and liking of classical music, as the estimated effect is much larger than the standard error.
- A larger absolute t-statistic indicates a stronger relationship between the independent variable and the dependent variable. It suggests that the estimated coefficient is more likely to be statistically significant.

P-values < 0.000001

- In this case, the very small p-value suggests that the observed relationship between education and liking of classical music is highly unlikely to be due to chance alone.

- There is a less than 0.0001% change that I would get a number as positive as 15.31 or as negative as -15.31 by chance, assuming the null is correct.

Sampling Distribution = probability distribution for a parameter

- The sampling distribution refers to the distribution of possible coefficient values that could be obtained from repeated sampling.
- Central Limit Theorem: as sample size increases, sampling distribution will approach normality, even if the population is not distributed normally
- The large t-statistic (15.331) indicates that the estimated effect of education on liking of classical music lies far in the tail of the sampling distribution. This means that the estimated effect is significantly different from zero, providing strong evidence for a real association between education and liking of classical music.

3-- Fully interpret the coefficient on Older (remember that “fully interpret” means that you consider the metrics that X and Y are measured in; you note other variables in the model; you note statistical significance; you make clear that you are predicting “on average”; and if it is an indicator variable, you note the reference or omitted category used). (12 points)

The coefficient on the variable "Older" represents the effect of being aged 50 or older on liking of classical music in Model 1.

Metric

- The variable "Older" is an indicator variable, also known as a dummy variable. It takes a value of 1 if the respondent is aged 50 or older and 0 otherwise. It is measured in terms of presence (1) or absence (0) of being in the reference category (those younger than 50).

Dependent Variable

- The dependent variable, Liking of Classical Music, is measured on a scale of 1 to 5, ranging from "very much do not like it" (1) to "very much like it" (5).

Other Variables in the Model

- Model 1 includes the independent variable Education, which measures the number of years of schooling.

On Average

- The coefficient on Older represents the average difference in liking classical music between individuals aged 50 or older (the reference category) and those younger than 50, while holding the education variable constant.

Statistical Significance

- The coefficient is statistically significant, as indicated by the t-statistic of 5.22 and the p-value of <0.000001. This suggests that the observed relationship between being aged 50 or older and liking of classical music is unlikely to have occurred by chance alone.

Interpretation:

- On average, individuals aged 50 or older tend to have a liking for classical music that is higher by 0.322 units compared to those who are younger than 50, while controlling for the effect of education.
- This means that age plays a significant role in predicting liking classical music, with older individuals showing a higher tendency to like classical music compared to their younger counterparts.

4-- (a) Interpret the R-squared in Table 2. What does an R-squared of 0.135 tell us about the overall fit of our model? (10 points). (b) Do you think that an R-squared of 0.135 is large or small? Justify your answer (10 points).

Interpretation of R-squared

- An R-squared of 0.135 means that approximately 13.5% of the variability in liking classical music is accounted for by the variables education and age in the model. In other words, these variables collectively explain 13.5% of the variation in individuals' liking for classical music.

Part 2

5-- In Table 4, fully interpret the logit coefficient on Age (remember that “fully interpret” means that you consider the metrics that X and Y are measured in; you note other variables in the model; you note statistical significance; you make clear that you are predicting “on average”; and if it is an indicator variable, you note the reference or omitted category used) (10 points)

Metrics and Other Variables:

- Independent Variable
    - Age: Respondent’s age (ranging from 18-98)

    - Education: Respondent's number of years of schooling (ranging from 0-20)
    - Uses_Facebook: Whether or not respondent uses Facebook (1=yes, 0=no) (**dummy variable**)
- Dependent Variable:
    - Uses_Instagram: Whether or not one uses Instagram (1=yes, 0=no) (**dummy variable**)

Other Variables in the Model:

- The model also includes the predictors Education and Uses_Facebook, which are controlled for when examining the relationship between Age and Instagram use.

Logit Coeff = -0.068:

- This coefficient measures the association between Age and the log-odds of using Instagram.
- On average, for every one-unit increase in Age, the log-odds of using Instagram decrease by 0.068, while holding Education and Uses_Facebook constant. This means that, as individuals' Age increases, their likelihood of using Instagram decreases, on average.

Statistical significance:

- The coefficient on Age is statistically significant, as indicated by the associated Z-statistic of -13.173 and the p-value of less than 0.000001. This suggests that the relationship between Age and Instagram use is unlikely to have occurred by chance.

6-- A) In Table 4, fully interpret the logit coefficient on Uses_Facebook (10 points). B) Now look at the Odds-Ratio column (the 3rd column in the table) – and look at the Odds-Ratio on Uses_Facebook and fully interpret that, too (remember that "fully interpret" means that you consider the metrics that X and Y are measured in; you note other variables in the model; you note statistical significance; you make clear that you are predicting "on average"; and if it is an indicator variable, you note the reference or omitted category used) (10 points).

A) Uses_Facebook Logit coefficient = 1.338

- This coefficient measures the association between Uses_Facebook and the log-odds of using Instagram.

- On average, individuals who use Facebook have a log-odds of using Instagram that is 1.338 units higher compared to individuals who do not use Facebook, holding other variables constant.

B) Uses_Facebook Odds-Ratio = 3.812

- Odds-Ratio represents the ratio of the odds, not the ratio of the probabilities
- On average, individuals who use Facebook have approximately 3.812 times higher odds of using Instagram compared to individuals who do not use Facebook, while controlling for the effects of Age and Education.
- In this case, the Odds-Ratio of 3.812 suggests that the odds of using Instagram are 3.812 times higher for Facebook users compared to non-users, on average.

Metric:

- Metrics of X (Uses_Facebook): 1 = Yes, 0 = No **dummy variable**
- Metrics of Y (Instagram Use): 1 = Yes, 0 = No **dummy variable**

Other Variables in the Model:

- The Odds-Ratio on Uses_Facebook represents the effect of Facebook usage on Instagram use while controlling for Age and Education.

Statistical significance:

- The Odds-Ratio on Uses_Facebook is statistically significant, as indicated by the p-value of less than 0.000001. This implies that the relationship between using Facebook and using Instagram is unlikely to have occurred by chance.