Instructions: Please read the following instructions thoroughly

- For the entire assignment, use `Python` for your analysis. Write your code in a Jupyter Notebook named as `[your-student-ID]_hw1.ipynb` (e.g., `2023-20000_hw1.ipynb`). The use of `R` is not allowed. You are allowed to use any libraries in `Python`.
- Type up your report and save as PDF named as `[your-student-ID]_hw1.pdf`. We do not allow the submission of a photo or a scanned copy of hand-written reports.
- Please upload two separate files on eTL, your report in PDF and the code in Jupyter Notebook. Submissions via email are not allowed. The violation of the filename or submission instruction will result in the penalty of 5 points. No worries if there are additional numbers at the end of the filename that appear when you submit more than once.
- You can discuss the assignment with your classmates but each student must write up his or her own solution and write their own code. Explicitly mention your classmate(s) you discussed with or reference you used (e.g., website, Github repo) if there is any. If we detect a copied code without reference, it will be treated as a serious violation of the student code of conduct.
- We will apply a grace period of late submissions with a delay of each hour increment being discounted by 5% after the deadline (i.e., 1-minute to 1-hour delay: 95% of the graded score, 1 to 2-hour delay: 90% of the graded score, 2 to 3-hour delay: 85%, so on). Hence, if you submit after 20 hours post-deadline, you will receive 0 points. No excuses for this policy, so please make sure to submit in time.

1. [**50 pts**] In this problem, you will use the `Carseats` data set attached in the assignment (`Carseats.csv`) for linear regression.

   (a) [5 pts] Fit a multiple linear regression model to predict `Sales` using `Price`, `Urban`, and `US`. Report the R^2 of the model.

   (b) [10 pts] Provide an interpretation of each coefficient in the model. Be careful — some of the variables in the model are qualitative!

   (c) [5 pts] Write out the model in equation form, being careful to handle the qualitative variables properly.

   (d) [5 pts] For which predictor variable j can you reject the null hypothesis $H_0 : \beta_j = 0$?

   (e) [5 pts] On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

   (f) [10 pts] How well do the models in Part (a) and Part (e) fit the data? Compare the two models. Can you say one model is better than the other based on R^2? Provide justification for your answer.

   (g) [10 pts] Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

2. [**50 pts**] In class, we used the example of the logistic regression model to predict the probability of `default` using `income` and `balance` on the `Default` data set attached in the assignment (`Default.csv`). In this problem, we will estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

   (a) [10 pts] Fit a logistic regression model that uses `income` and `balance` to predict `default` using the validation set approach, and estimate the test error of this model. In order to do this, you must perform the following steps:

      i Split the sample set into a training set and a validation set.

      ii Fit a multiple logistic regression model using only the training observations.

      iii Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the `default` category if the posterior probability is greater than 0.5.

      iv Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

   (b) [10 pts] Provide an interpretation of each coefficient in the trained model.

   (c) [15 pts] Perform 5-fold cross-validation using the same model as in Part (a). Write your own code for the K-fold cross-validation. How does the validation error of the cross-validation the differ from to the results in Part (a)?

   (d) [15 pts] Now consider a logistic regression model that predicts the probability of `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the 5-fold cross-validation set approach. Comment on whether or not including a dummy variable for student would lead to a reduction in the test error rate.