CSE 2600: Introduction to Data Science and Engineering
Assignment 3

Instructions:

- You must support your answers to receive credit.
- Answers can be typed or handwritten, and should be readable.
- Submit the assignment in one file via PDF on HuskyCT.

1. (6 points) Adapted from ISLP 4.8.6.

   Suppose we collect data for a group of applicants to a data science bootcamp with variables:

   - X_1 = coding practice hours completed,
   - X_2 = prior programming experience (in years), and
   - Y = admitted to the bootcamp.

   We fit a logistic regression and produce estimated coefficients

   $$\beta_0 = -4.2, \beta_1 = 0.04, \text{ and } \beta_2 = 0.8.$$

   (a) Estimate the probability that an applicant who has completed 60 hours of coding practice and has 2 years of programming experience is admitted to the bootcamp.

   (b) How many practice hours would the applicant in part (a) need to study to have a 90% chance of being admitted?

   (c) Describe in words the meaning of "$\beta_0 = -4.2$" in the context of this model.

2. (16 points) Adapted from ISLP 4.8.13.

   Load the `OJ` data set, which is part of the ISLP library. Details can be found at `https://intro-stat-learning.github.io/ISLP/datasets/OJ.html`.

   As a first step, create two subsets of this data:

   - `Train`, containing only records with `StoreID` 1, 2, 3, or 4, and
   - `Test`, containing remaining records with `StoreID` 7.

   We will use these subsets to predict Citrus Hill purchases.

   (a) In `Train`, locate three variables: Citrus Hill loyalty, indication of a Citrus Hill special, and price difference between the brands. Calculate the mean, standard deviation, median, minimum value, and maximum value for each of these variables.

   (b) Provide a correlation matrix for the three variables in (a).

   (c) Use the `Train` subset to perform a logistic regression with `Purchase` as the response and the three variables from (a) as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

   (d) Using `Train` and a probability threshold of 0.5, compute the confusion matrix and overall fraction of correct predictions. Also note the percent of false positives and the percent of false negatives.

(e) Now refit the logistic regression model on the `Train` data with only the *significant* predictors. For this model, use `Test` to compute the confusion matrix, fraction of correct predictions, percent of false positives, and percent of false negatives.

(f) Use Naive Bayes on the `Train` data with only the significant predictors from (e). Use `Test` to compute the confusion matrix, fraction of correct predictions, percent of false positives, and percent of false negatives.

(g) Build three k-Nearest Neighbors models on the `Train` data, using $k = 5, 50, 150$, and only the significant predictors from (e). For the $k = 50$ model, use `Test` to compute the confusion matrix, fraction of correct predictions, percent of false positives, and percent of false negatives.

(h) Provide a ROC curve for each of the five models in parts (e), (f), and (g) on the same plot. Which model yields the highest AUC?

3. (12 points) Adapted from ISLP 12.6.3.

In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as follows:

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Plot the observations using a single color for all data points.

(b) Randomly assign a cluster label to each observation. You can use the `np.random.choice()` function to do this. Report the cluster labels for each observation.

(c) Compute the centroid for each cluster.

(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

(e) Repeat (c) and (d) until the answers obtained stop changing.

(f) In your plot from (a), color the observations according to the cluster labels obtained.

4. (16 points) Adapted from ISLP 12.6.9.

Load the `CEV2021` data set, which is available on HuskyCT. Source data and details can be found at `https://data.americorps.gov` (search “CEV findings: state”).

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Report your answer with a dendrogram.

(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

(c) Cut the dendrogram at a height that results in four distinct clusters. Which states belong to which clusters?

(d) Use K-means clustering and $K = 3$ to obtain new clusters of states. Which states belong to which clusters?

(e) Repeat part (a), but first rescale and replace your variables so that they have mean 0 and standard deviation 1 (i.e., perform the clustering on the z-scores of the features, rather than the features themselves). Provide the new dendrogram.

(f) Cut the new dendrogram from part (e) at a height that results in three distinct clusters. Which states belong to which clusters?

(g) Using the z-scores created in part (e), use K-means clustering and $K = 3$ to obtain new clusters of states. Which states belong to which clusters?

(h) Create a scatter plot of the states with axes `Voting_Local` and `Organization_Membership` which assigns a color to each point based on the clusters found in part (g).