Problem 1

*Adapted from *Predicting Life Expectancy in the United States* – page 394 in The Analytics Edge by Bertsimas, O'Hair, and Pulleyblank.

Explaining Life Expectancy in the United States during the 1970s

This problem will focus on explaining life expectancy in the United States during the 1970s. The file *StateData.csv* includes data collected during the 1970s for all fifty US states. The data dictionary is given in Table 1.

Table 1: Data dictionary for StateData.csv

Variable	Description
Population	Population
Income	Per capita income
Illiteracy	Illiteracy rates (percentage of state population)
LifeExp	Life expectancy (in years)
Murder	Murder and non-negligent manslaughter rate per 100,000 population
HighSchoolGrad	High-school graduation rate
Frost	Average number of days (over the last 30 years) with a minimum temperature below freezing in the state capital or a major city
Area	Land area (square miles)
Longitude	Longitude of the center of the state
Latitude	Latitude of the center of the state
Region	The region that the state belongs to (Northeast, South, North Central, or West)

1. (a) Create a scatterplot of all the states center locations using the Longitude and Latitude (make sure you have them on the correct axis!).
   (b) According to this plot, what is the most southern state? Is that correct? Which state is actually the most southern?

2. (a) On average, what state has the least number of days below freezing? How many?
   (b) Which region of the United States (West, North Central, South, or Northeast) has the highest average high school graduation rate?
   (c) What is the high school graduation rate in each region weighted by population?

3. Create a box plot of the murder variable for each region (i.e., there should be four box plots in the figure - you can put them all on the same axes).
   (a) Which region has the highest median murder rate?
   (b) Which region has the largest range (difference between max and min)?

4. Fit a linear regression model to explain life expectancy using seven features: Population, Income, Illiteracy, Murder, HighSchoolGrad, Frost, and Area. Use the entire dataset to fit your model and be sure to include an intercept term.

   (a) What is the regression equation?

   (b) What is the interpretation of the coefficient for Income?

   (c) Create a scatterplot with Income on the x-axis and LifeExp on the y-axis. Does this relationship agree with the coefficient? Why or why not?

   (d) What is the interpretation of the coefficient for Murder?

5. Fit a linear regression model using only 3 features: Murder, HighschoolGrad, and Frost.

   (a) What can you say about cold weather and life expectancy? Do you agree? Why or why not?

   (b) What is the range (difference between max and min) of the Frost variable? How many years (LifeExp) does that equate to?

Problem 2

Predicting invasive species

This problem will focus on using logistic regression to predict the likelihood that an invasive tree species is present on a particular plot of land in the forest. The file *SpeciesData.csv* includes a large data set with 11,684 observations and 54 features. Our target (column name *Target*) is a binary variable that indicates whether or not the invasive species is present. The features are shown in Table 2

Table 2: Data dictionary for SpeciesData.csv

Variable	Description
Elevation	Vertical elevation from sea level (meters)
Aspect	The direction that the plot of land is facing (degrees azimuth)
Slope	The average slope of the plot of land (degrees)
HdistWater	Horizontal distance to nearest surface water (meters)
VdistWater	Vertical distance to nearest surface water (meters)
HdistRoad	Horizontal distance to nearest road (meters)
Shade9	Hill-shade index at 9am during summer solstice (0-255)
Shade12	Hill-shade index at 12pm during summer solstice (0-255)
Shade3	Hill-shade index at 3pm during summer solstice (0-255)
HdistFire	Horizontal distance to nearest wildfire ignition point (meters)
WA*	4 binary columns represent the wilderness area designation (Rawah, Neota, Comanche, Cache)
Soil*	40 binary columns represent the soil type (1-40)

Your first task is to train a logistic regression with L1-regularization and a logistic regression with L2-regularization.

1. For each model (L1 and L2), search over 10 lambda values in the logarithmic scale between 0.0001 and 10000 using 10-fold cross validation and AUC as your performance metric. Before training your model, make sure you scale your data to the interval $[0, 1]$.

   (a) Which model (L1 or L2) took longer to train? Why?

   (b) Create a boxplot figure for each model (L1 and L2). Each figure should include 10 different boxplots of AUC, corresponding to the 10 different lambda values.

   (c) What is the best lambda value for L1-regularization? L2-regularization?

2. Randomly split the dataset using a ratio of 70/30 for training/testing. Use the training set to fit one L1 and one L2 model using the corresponding best lambda value from part 1. Predict the probability of invasive species using the test set.

   (a) Plot the ROC curve for each model. Which model do you prefer? Why?

   (b) Suppose we want a true positive rate of at least 0.85. What is the corresponding false positive rate for each model? What is the corresponding threshold rule for each model?

   (c) For each model, use the threshold rule from 2b to convert your predicted probabilities to binary classes and create a confusion matrix. Which model do you prefer? Why?

Python hints:

- For part 1, you may find the following functions helpful: *MinMaxScaler* and *LogisticRegressionCV*
- For part 2, you may find the following functions helpful: *train_test_split*, *LogisticRegression*, *roc_curve*, and *roc_auc_score*
- You need to manually specify the scoring function for *LogisticRegressionCV* (in python AUC is referenced by the string *roc_auc*)
- Use the *liblinear* solver with a maximum of 100 iterations for L1-regularization
- Use the *lbfgs* solver with a maximum of 1000 iterations for L2-regularization.

Problem 3

Consider a binary classification explanation problem.

1. If we model the problem using linear regression, what assumption(s) will be violated? Why?

2. Describe the linearity assumption for logistic regression (i.e., what must be a linear combination of the features?).

3. Provide two assumptions of logistic regression that are identical to linear regression.