In this analysis we are supposed to use logistic regression on a dataset so that we can make a machine learning model that can predict the classification of the outcome labels based on the input data from the dataset. On Kaggle I found an interesting dataset that I wanted to use the logistic regression on. It is a gender classification dataset which has some characteristics of facial features which would be used to see if they can be used to identify if that person is male or female. I found this dataset to be interesting because in today's trends there are many guys who look effeminate while there are a lot of girls out there who look more masculine. So, with this dataset I wanted to see if it is possible to correctly identify the genders of people based on the characteristics that are provided in the dataset.

First, I imported the necessary libraries to do this analysis. Then, I read in the dataset as a data frame on Google Colab. I took a look at what the dataset looks like with shape, describe, info, and dtype functions. After that, I created a count plot for the male to female ratio and found that it is a 50/50. Then, I made count plots for gender with the hue based on the characteristics in the dataset. I found that male and females tend to have longer hair, males tend to have wider and longer noses, males tend to have thin lips, and males have a long distance between nose and lips. After these visuals, I checked to see if the dataset has any missing values. Finding that there are none, I used a distribution plot to visualize the distribution of forehead height, and forehead width in the dataset. For these two characteristics, they tend to be on the lower side of the spectrum. Finally, I made a data frame out of the y_predicted results and named it "Predicted Gender". I made a new data frame that combined the original dataset with this additional "Predicted Gender" column, to show that results of the regression side by side with the actual data.

To begin the logistic regression, I one hot coded the gender to 0's and 1's (1 meaning male and 0 female). I then replaced the original gender column (which were strings) with this numerical column. After that, I created a variable "X" and gave it the characteristics used to determine gender. I gave the "y" variable the label which is the new gender column. Next, I imported train_test_split from sklearn to split the dataset into 70% training, and 30% testing with a random_state of 10 to make the results repeatable. Then, I imported the logistic regression model from sklearn and passed that to the variable "lr". Then, I fitted the X_train and y_train data into "lr" and made it predict the y_predicted from the X_test data. I then generated a confusion matrix from this model and found this model to be very accurate. True positive is 725, false positive of 15, false negative of 24, and true negative of 737. From the classification report generated, all the scores are pretty high ranging from accuracy, precision, recall, and f1 score. They are all above .95, which states that this model is good at classifying the genders of the participants in this data set. This also shows that based on the characteristics of long hair, having a wide nose, having a long nose, forehead height, forehead width, thin lips and, distance from nose to lips, we can accurately tell if someone is a male or female.