STAT 477
Take Home Final Exam

Directions: Below is the take home exam for the material covered in Module 3 - Sections 1 and 2. You will not need material outside of these two sections. You may only receive help from me on the take home exam. Receiving assistance from any other source will be considered academic dishonesty and treated as such.

Your responses to the questions below should be prepared on the computer and uploaded to Canvas before the deadline. Your responses should be written in paragraph form. You should include all tables and graphs as a part of this paragraph structure in your write-up, not separately at the end.

Assignment: The purpose of the study described below is to compare residential intervention programs designed to reduce drug abuse and prevent behavior associated with a high risk of HIV infection. A residential invention program was carried out in a community in which the target subjects lived. The two intervention programs compared in this study are referred to as program A and program B. Each program had a short duration version and a long duration version and it was of interest to also determine if the duration of the program had any impact on the effectiveness of the program in preventing high risk HIV behavior related to drug abuse.

A total of 300 subjects participated in program A and these subjects were randomized to receive either the short (3-month) or long (6-month) version of this program which incorporated elements of health education and drug use relapse prevention. Participants in this program were taught to recognize high-risk situations that are triggers to relapse and were taught skills to enable them to cope with these situations without using drugs.

A total of 125 subjects participated in program B and these subjects were randomized to receive either the short (6-month) or long (12-month) version of this program which required subjects to follow a highly structured lifestyle in a communal living setting.

The outcome variable for this study is whether the subject remained drug free for one year after completing the treatment program.

Since the researchers were not able to randomly assign participants to the particular program, they also collected other demographic information about each participant to use in the study. These variables are the focus of this assignment.
The data file **impact.csv** contains data on the 425 subjects in this study. Here is a description of each variable in the data file.

Variable	Description
ID	Subject identification number (three digit code)
AGE	Age at enrollment into the study (recorded in years)
BECK	Beck depression score (higher values indicate more severe depression)
IVHX	IV drug use history at enrollment into the study (Never, Previous, Recent)
NDTX	Number of times subject previously treated for drug use (coded 0 to 40)
RACE	Subject's race (White, Other)
DUR	Duration of the program (Short, Long)
PROG	Intervention program (A, B)
DFREE	Response variable (1 = DrugFree, 0 = Otherwise)

Find the best logistic regression model with a collection of background variables: AGE, BECK, IVHX, NDTX, and RACE to predict the probability of staying drug free (DFREE) for one year after completing the program. Answer the following questions about your model.

1. What are the explanatory variables in your final model? Describe the process by which you chose this model.
2. Does your model as a whole help to explain the probability of staying drug free in this population? Report the appropriate test statistic and p-value to support your answer.
3. Do each of your explanatory variables help to explain the probability of staying drug free in this population given the other explanatory variables are in the model? Report appropriate test statistics and p-values to support your answers.
4. Provide the confusion table and associated statistics for your model. Write a summary of these statistics.
5. Provide the ROC curve for your model. Give the area under the ROC curve and write an interpretation of its value.