Linear Regression Assignment 2
Big Data Science

This assignment should be done in your groups. Feel free to use Python or R for this assignment.

1. The data in icudata.csv are 200 observations from a much larger study on survival of patients admitted to the ICU.
    a. Fit a logistic regression model that uses age (AGE), race (RACE), whether or not CPR was administered (CPR, 0 for no, 1 for had been administered), systolic blood pressure at admission (SYS), heart rate at admission (HRA), and type of admission (TYP). Do not do any parameter selection. Create a table with the coefficients from the regression.
    b. Explain the effect CPR has on survival. How much does receiving CPR increase your likelihood to survive?
    c. Now, using the same variables fit a model using the lasso technique. Using cross-validation, what is the optimal alpha value?
    d. Using the alpha value and model from part (c), what are the coefficients for each of the parameters that were originally put in the model?
2. TED_talk_data.csv contains data for 2,550 TED talks. We are interested in finding a way to predict the number of views a talk will get.
    a. Using the tag column, create a new column for each unique tag called TAG_xxx where xxx is the name of the tag in the tag column. The value will be TRUE or FALSE based on if the tag is contained in that talk's tag column. For example, if the value for the tag column is ['children', 'creativity'] in row 1, there will be a new column TAG_children with TRUE, TAG_creativity with TRUE, and all other TAG_xxx columns will be false.
    b. Using the ratings column, create a new column for each rating category (14 in total). The value will be the count for the associated category for each row. For example, if the value is [{'id': 7, 'name': 'Funny', 'count': 19645}, {'id': 1, 'name': 'Beautiful', 'count': 4573}]. Then the RATINGS_Funny column will be 19645 and RATINGS_Beautiful column will be 4573.
    c. Using LASSO, fit a model using comments, duration, number of speakers (num_speaker), the tag data (TAGS_xxx), and the ratings data (RATINGS_xxx). The TAG_ columns are true and false based on if the

talk was tagged with the specified tag. The RATINGS_ columns the number of "votes" for the rating specified.
Note: There are additional columns in the dataset, ignore these.

d. Using Cross-Validation, what is the optimal value for lambda?
e. What are the top 10 best and 10 worst topics (tags) for gaining viewership? What are the coefficients for these topics?
f. What is the least important characteristic (rating) of a TED talk in terms of getting more views?