1. When are we going to get invited to box folder?
2. Do we only analysis on the lists? => yes, but I can do more.
3. When do R coding, do we separate function and scripts like what 133 did?
4. More information of how to write and how much to write a report?
5. Do we need to divide <https://archive.ics.uci.edu/ml/datasets/Census+Income> data into training and test set by ourselves?

Or, are we allowed to use training and testing set given in the github? => we can use the one there.

1. Is there any restriction in the project? For example, using embedded functions or so..?
2. Do we need to make slides?
3. So, when you said “model selection,” does it mean that which are the best out of classification tree, bagged tree, and random forest?
4. How are you going to grade this? Is there any rubric?
5. Do we do EDA on both training and testing?
6. So our goal is reduce the test prediction error. Do we have to reduce error only with the given methods, or are we allowed to use any method to achieve the goal?
7. Why do we really care about OOB error rate if this is basically the error rate from training set…??
8. I am confused with the concept of fit a tree to the residual from the model in boosting tree. What do you mean by residual here?
9. APM introduces the way how to deal with categorical variable on pg 373, but I am not sure whether I understand it correctly. Do you recommend me to split into dummy variable before modeling?
10. APM pg 381 says that when there is a missing value, we need to weight it differently. Can you explain it to me in detail how we should deal with missing values…?
11. Are we learning AdaBoosting or simple gradient boosting? They look quite similar (pg 390 and 391)
12. Need to check my missing value imputation….
13. Do we find outliers for categorical variables?
14. And, I am not sure whether I am understanding fnlwgt variable correctly. Is it possible to have huge number for fnlwgt? What is the boundary for fnlwgt?
15. We do not dummify income variable. Right?
16. When we do dummify and discretize, we apply them both on training and testing sets. Right?
17. How many intervals to make when discretizing?
18. Do we need to dummify the discretized variables as well?
19. When we do dummify, since there is no intercept we need to do one hot encoding, then does it mean that we always have one column less than the level of the variable. Then, will it be hard to interpret it after we do tree as there are so many dummified variables?
20. What do you mean by “**Association between each predictor and the response”** in EDA?