Consider the `Caravan.csv` data set. It is of interest to predict `Purchase`. Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations.

1. Use 2 pivot tables to show that (in the original data set) the variables `PVRAAUT` and `AVRAAUT` are highly unbalanced (having most rows belonging to a few categories). Remove these variables from the dataset.
2. Fit a random forest model with 500 trees and `max_features = 29` to the training set with `Purchase` as the response and the other variables as predictors. Use `random_state = 1`. What predictor appears to be the most important? Report the test accuracy rate.
3. Fit a boosting model to the training set with `max_depth` = 4 and `Purchase` as the response and the other variables as predictors. Use 1000 trees, learning rate 0.01, and `random_state = 1`. What predictor appears to be the most important? Report the test accuracy rate.
4. Report the test accuracy rate when KNN is used to predict `Purchase` using 5-fold cross validation to find the best number of neighbors.
5. Find the test accuracy rate when logistic regression is used to predict `Purchase.`

Submit your report as a pdf file onto Blackboard. Be aware that your pdf must

- show your name and USC ID.
- is made of letter size pages in portrait format (not landscape).
- show Python commands fully displayed and not truncated.