Neha Maddali

Problem 1:

a. Neha Maddali
   Taylor Turner
   Jung Ho Suh
   Kordell Schrock
b. Neighborhood Watch
c. Data set: https://data.iowa.gov/Correctional-System/Current-Iowa-Correctional-System-Prison-Population/xbcv-c6t2
d. We'll have everyone work on exploratory analysis and then split our two questions amongst ourselves (two people work on one question and two people work on the other)

Problem 2:

a. I do not think this is true. M_{k+1} does not need to have a subset of the predictors that are found in M_k. The "best" model of each size is determined you compare the RSS values to all models of the same size. Therefore, this statement is not true.
b. This applies to forward selection. M_{k+1} comes from increasing the predictors in M_k
c. $M_{k,\ subset}$ should have the smallest training MSE since best subset selection in the only method where all the predictors of all models are considered.
d. $M_{k,\ subset}$ should have the smallest test MSE since best subset selection considers more models than the other techniques we've covered.

Problem 3:

a. I implemented forward and backward model selection on the entire dataset. This is because you do not need to split the data for AIC and BIC.

```
regfit.fwd = regsubsets(Apps~.,data=College,nvmax=17, method="forward")
regfit.bwd = regsubsets(Apps~.,data=College,nvmax=17, method="backward")
```

b. Forward Selection BIC: model 10

App = -100.51668243 + -575.07060789*PrivateYes + 1.58421887*Accept + -0.56220848*Enroll + 49.13908916*Top10perc + -13.86531103*Top25perc + -0.09466457*Outstate + 0.16373674*Room.Board + -10.01608705*PhD + 0.07273776*Expend + 7.33268904*Grad.Rate

Forward Selection AIC: model 12

App = -157.28685883 + -511.78760196*PrivateYes + 1.58691470*Accept + -0.88265385*Enroll + 50.41131660*Top10perc + -14.74735373*Top25perc + 0.05945481*F.undergrad + 0.04593068*P.undergrad + -0.09017643*Outstate + 0.14776586*Room.Board + -10.70502848*PhD + 0.07246655*Expend + 8.63961002*Grad.Rate

Backward Selection BIC: model 10

App = -100.51668243 + -575.07060789*PrivateYes + 1.58421887*Accept + -0.56220848*Enroll + 49.13908916*Top10perc + -13.86531103*Top25perc + -0.09466457*Outstate + 0.16373674*Room.Board + -10.01608705*PhD + 0.07273776*Expend + 7.33268904*Grad.Rate

Backward Selection AIC: model 12

App = -157.28685883 + -511.78760196*PrivateYes + 1.58691470*Accept + -0.88265385*Enroll + 50.41131660*Top10perc + -14.74735373*Top25perc + 0.05945481*F.Undergrad + 0.04593068*P.Undergrad + -0.09017643*Outstate + 0.14776586*Room.Board + -10.70502848*PhD + 0.07246655*Expend + 8.63961002*Grad.Rate

The AIC and BIC for both forward and backward are the same models. Both BICs were model 10. Both AICs were model 12. The predictors in model 10 are a subset of model 12 however they do not contain the two predictors F.Undergrad and P.Undergrad from model 12.

c. AIC and BIC can be done on the full dataset. The AIC and BIC were the same as in part b. AIC was model 12 and BIC was model 10.

AIC model 12:

App = -157.28685883 + -511.78760196*PrivateYes + 1.58691470*Accept + -0.88265385*Enroll + 50.41131660*Top10perc + -14.74735373*Top25perc + 0.05945481*F.Undergrad + 0.04593068*P.Undergrad + -0.09017643*Outstate + 0.14776586*Room.Board + -10.70502848*PhD + 0.07246655*Expend + 8.63961002*Grad.Rate

BIC model 10:

App = -100.51668243 + -575.07060789*PrivateYes + 1.58421887*Accept + -0.56220848*Enroll + 49.13908916*Top10perc + -13.86531103*Top25perc + -0.09466457*Outstate + 0.16373674*Room.Board + -10.01608705*PhD + 0.07273776*Expend + 7.33268904*Grad.Rate

d. To report the model with smallest test MSE, you need to implement this on the training set.

Model 5 had the smallest test MSE.

Model 5: App = -514.25299848 + 1.59096970*Accept + -0.51519696*Enroll + 35.54981407*Top10perc + -0.10394174*Outstate + 0.08515977*Expend

e. The model with the smallest test MSE for a 60/40 split was model 3

Model 3: App = -907.12390883 + 1.39137076*Accept + 22.74759314*Top10perc + 0.04350431*Expend

One major advantage of forward selection is that with AIC or BIC as the criteria, we don't have to know what the best split of the dataset is. There would also not be any randomness involved when implementing these techniques on the full dataset.