Gina O'Riordan
Behavioral Data Science - East
Final Project 12.1.18

After a browse through each of the data sets provided, many were intriguing but one stood out as most interesting to me for this project. The Texas Execution data reflects information from the Texas Department of Criminal Justice regarding offenders who were convicted and executed in the state, as well as some data on their victims. Since there was such a rich amount of data about the offenders and their crimes, time spent on death row, final statements, and victims, I thought this would be a great set of data from which to glean insights. What I was most interested in were the factors that influenced three variables: whether the offender had a prior record, their age at the date of offense, and the amount of years on death row. By understanding the influences on an offender's prior record, the Department of Criminal Justice might be able to work with the government to create programs to prevent repeat offenders by tailoring to certain subgroups of the criminal population. Knowing the executed offender's age at the date of offense and what factors contributed to a younger or older person committing a crime could also help the government work to prevent crimes by age group. Finally by analyzing the factors contributing to the years on death row, the Department of Criminal Justice might be able to better organize logistics of sentencing and jailing these criminals to best use taxpayer dollars.

At the beginning of the semester, I had numerous hypotheses that were not general enough to achieve the goals mentioned above. I reflected on what insights I hoped to glean from the project and discussed my thoughts with the Professor before narrowing down my hypothesis to the following: I am interested in what factors predict a prior record, age of offense, and years on death row of an executed Texas inmate using the following variables as predictors: countyorCountry, educationYears, codefsYes, totalVictims, femaleVictim, foreignNational, race2. Therefore, I hoped to understand if an inmate's home town being rural or urban, their years of education, if they had codefendants, the total number of victims, if they had a female victim, if they were a foreign national, and if their race predicted the three main variables.

My execution plan included an exploratory data analysis. There were 518 total observations in the dataset; not a huge number, but enough to gain some insights. I only used a portion of the original dataset variables, so I had to first pare it down to just the 9 I needed plus an identifying variable executionNumber. This variable was not analyzed since it was just an identifier that would eventually be removed from the dataset before running analyses. Then I checked for missing values, finding only the countyorCountry variable to be a bit sparse but did not have zero variance so it was ok to use in my analysis (more details below).

I looked at the means and standard deviations of the numeric variables. It seemed the offenders' average age was relative low at 26.4 but there is a large standard deviation of almost 8 years, which means the offenders' ages are probably pretty varied. This might require additional analysis after to breakdown insights by age groups. The average time spent on death row was 11.1 years but with a smaller deviation than age at 3.9, meaning most inmates spent around a decade waiting for execution. If I'm able to pull valuable insights on this variable, it could majorly affect logistics by the Department of Criminal Justice if inmates are on death row for so long a period of time. The mean years of education for offenders is 10 years with a 2 year standard deviation, putting most offenders with a high school education or lower. This will make insights easy to review with most inmates in the same general tier of final education. And the total victim average is just above 1 with a standard deviation less than 1, so there's minimal variation between how many victims each executed prison had. Therefore, the insights gleaned from analysis will probably only be relevant for offenders with minimal victims (in other words non-serial offenders).

From visualizing these numeric variables, there were no new insights that appeared in the histograms; however they did reinforce the information gleaned from the previous EDA steps. For example, the ageatDateofOffense variable is right skewed, reflecting the 26.4 year mean age (Appendix 1). The yearsonDeathRow is also right skewed with a mean at around 10 years (Appendix 2). The totalVictims is severely right skewed, with almost all executed

prisoners with one victim, and only a few with 2-6 victims and one with 12 (serial) (Appendix 4). Finally, the educationYears variable is left skewed, with the majority of executed between a middle school and high school education, as indicated by the mean of 10 years and 2 year standard deviation (Appendix 3).

Finally, I looked at the distribution tables for the factor variables. The priorRecordYes and codefsYes variables were quite evenly distributed, while the race2 variable was heavily leaning towards black and white races. This uneven distribution for the race data will be kept in mind while running the analysis. The countyorCountry variable presented a challenge, however; it listed the county names from whence the executed inmates came, but I was mostly looking for rural v. urban data. Since the counties are listed as factors, this variable needed to be turned into a binary variable. Based on geographic research done online, I looked at the top 10 metropolitan areas (in order of decreasing size) and matched them to the following counties (major cities in parenthesis): Dallas (Dallas), Tarrant (Fort Worth), Harris (Houston), Bexar (San Antonio), Travis (Austin), El Paso (El Paso), Nueces (Corpus Christi), Bell (Temple). The counties for the city of Mission, Brownsville, and Beaumont did not appear in the dataset and were thus not included in the pre-processing. By using an if-else statement to turn the county names listed here to be 'yes' and all other observations to be 'no,' I turned the variable into a factored binary variable called metroArea. Then I proceeded to remove the original countyorCountry variable. Since this variable does still have a majority of missing values, I will be sure to use it with precaution in my analyses. Last but not least, the data set was complete, so I wanted to check for zero variance; no variables had zero variance and thus I was ready to move onto the analysis.

Since I'm trying to determine the factors that affect the main three variables, I decided to use decision trees and random forests as my algorithms of choice. I first made a data set without the exeuctionNumber identifier variable and created the test and training data with a 20-80 split before setting the seed at 1842. For each variable (priorRecord, ageatDateofOffense, and yearsonDeathRow), I did five algorithms: a decision tree with all remaining variables as predictors, then a pruned tree with the top predictor variables from the first tree, repeated these two steps with random forests made of 100 trees, and finally did a random forest with the same number of tree repetitions but using the variable predictors from the pruned decision tree. After analyzing the top predictor variables and resulting metrics, I pulled together my insights.

The priorRecord analysis found the educationYears, ageatDateofOffense, and yearsonDeathRow variables to be important predictors in the first tree (Appendix 5). However, it didn't make much sense that the amount of years on death row indicated whether someone had a prior record or not, so I only used the first two variables when building the pruned tree. I saw the root node error increase slightly in this pruned tree (by 0.2%), but the majority of metrics remained the same. When I built the initial random forest, I found ageatDateofOffense, yearsonDeathRow, educationYears, and race2 as important predictors. Again, I removed yearsonDeathRow and added the other three variables to the formula in the second random forest. In this algorithm, classification errors decreased as well as the out of bag errors from 36.4 to 35%. The final random forest using the decision tree two variables saw the out of bag error rate go up slightly to 37.44%.

The ageatDateofOffense analysis had many breaks in the initial tree including on the race2, priorRecordYes, and educationYears variables (Appendix 6). Using those in the pruned tree, it became a much cleaner tree with less breaks, but the root node error stayed exactly the same and there was a relatively higher error (by 8%) on the final split in the second tree than the first. The initial random forest saw yearsonDeathRow, educationYears, race2, and priorRecordYes as the most important predictors. However, the yearsonDeathRow, similar to pruned tree for priorRecord, did not make sense as a predictor for the age variable and was thus not included in the second random forest. The mean of squared residuals went up slightly (by 0.28) and the percentage of variance explained increased by just under 1% in the final

random forest. In the final random forest with the pruned decision tree variables, the mean of squared residuals dropped even further down by 7 to 44.34, with a jump to 14.53% variance explained.

The yearsonDeathRow analysis had minimal breaks on the ageatDateofOffense, codefsYes, and femaleVictim variables (Appendix 7). These were inputed as the only predictors in the pruned tree, which saw the root node error stay exactly the same, but there was a lower relative error not he final break (by 6%). The initial random forest saw the ageatDateofOffense, educationYears, and race2 variables as most important, interestingly quite different from the decision trees. When these three were used in the second random forest, the percent variance explained actually went down to -23.76. Since this was so much lower than the -5.65% of the original random forest, I changed the formula to predict on the three variables from the decision tree and ended up with a much better -7.49 percent of variance explained (thought not great overall).

The results of these algorithms did address the research questions. The inmate's age at the time of the offense and years of education dictate their prior record status. The inmate's age at the time of the offense in turn is dictated by their race, prior record status, and education years. These first two sets of results matched some of my assumptions, but the amount of cross over between predictor variables in the first two analyses surprised me. It seems that those few variables of age, education, and prior record are all very much related to one another. I would very much like to do a further analysis using different methods to find out more about these relationships. If these connections are defined even more, the Texas Department of Criminal Justice can work closely with government groups and legislators to help curb crimes for certain age groups, education levels, and prior offenders.

The inmate's years on death row are dictated by their age at the time of the offense, number of codefendants, and whether or not their victim(s) was female. This final result definitely matched my assumption; criminals on death row who were older (usually with a longer criminal timeline) are executed quickly, as well as those who victimized women. I would like to explore this data a bit more, because I'm curious if the number of codefendants being high or low dictates a longer time on death row; I'm not familiar enough with the current justice system in Texas to make an assumption either way.

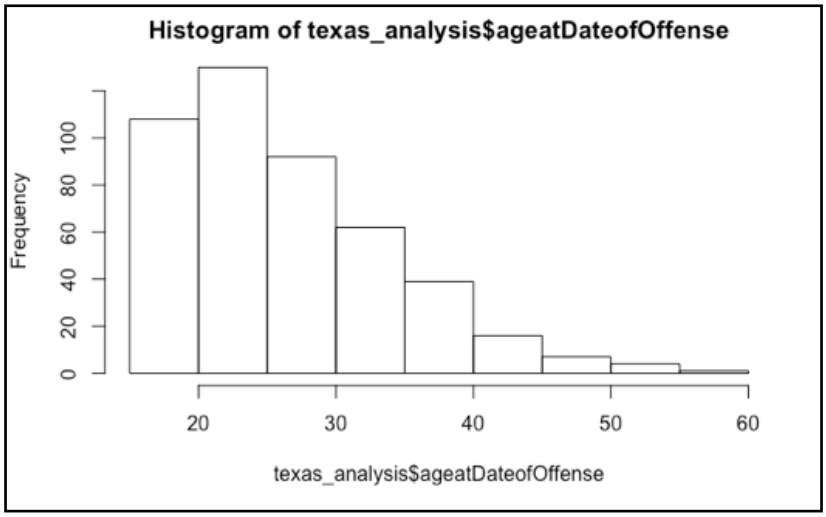
Decision trees and random forests are great starting points to analyses. They help sift through the total amount of variables in a dataset to find the ones that truly impact the response variable. However, they do just that, show a quick picture rather than a true deep dive into the data. If I were to repeat these exercises with a full data set, more observations, or more trees in the random forests, i may get slightly more accurate data (lower error rates, etc.) but would not have enough information to make clearer insights than I have right now. That would require different types of algorithms and more time! If I had such time, I would definitely inquire further into location information, as I do think there is a connection between prior record and age at date of offense and whether the inmate was from the city or the rural areas of a state. I'd also like to have more segmented age data to help find insights into specific age groups rather than just the overall execution data. I would also love to have more data to work with; though 518 observations is quite a bit, having something like 5000 observations would make a much stronger analysis and impactful insights with supporting data.

Overall, I think this analysis could provide the Texas Department of Criminal Justice useful information to both help prevent crimes that warrant the death penalty and to help the penal system reallocate funds and resources based on how long inmates are on death row.

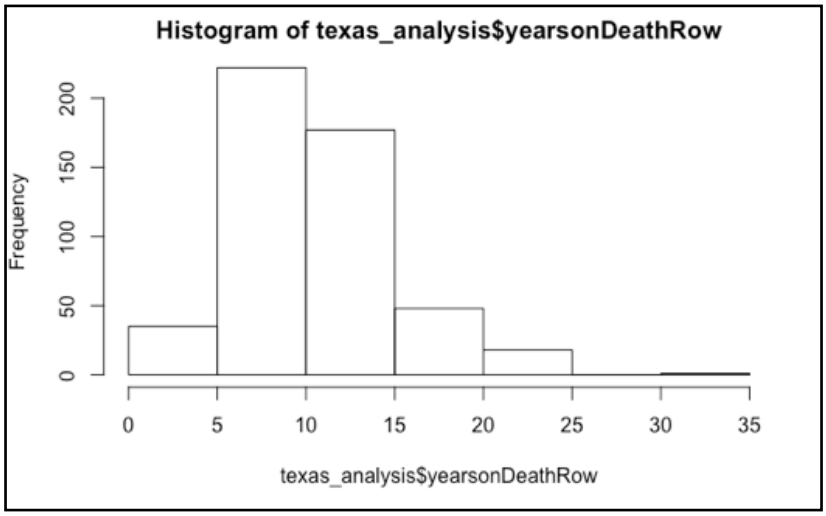
*NOTE: All data outputs can be found in the 'BDS Project.Rmd' file in the zip folder.

Gina O'Riordan
Behavioral Data Science - East
Final Project 12.1.18

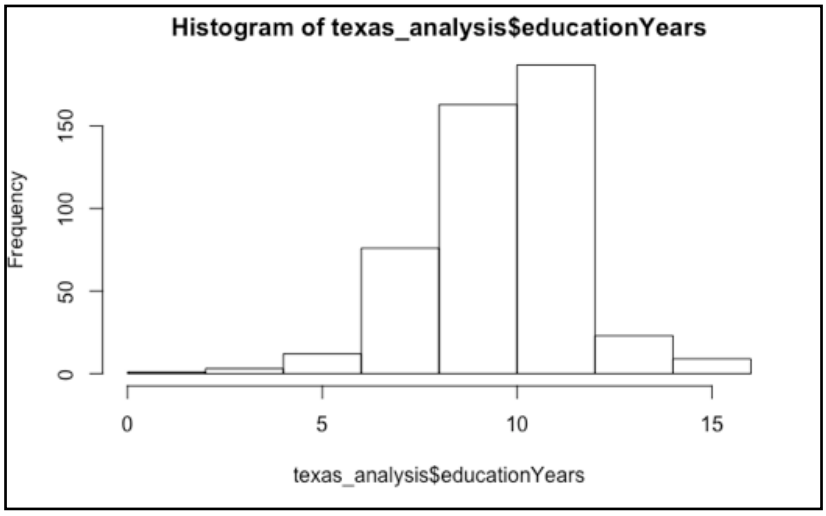
Appendix 1: EDA Histogram of 'ageatDateofOffense' Variable



Appendix 2: EDA Histogram of 'yearsonDeathRow' Variable

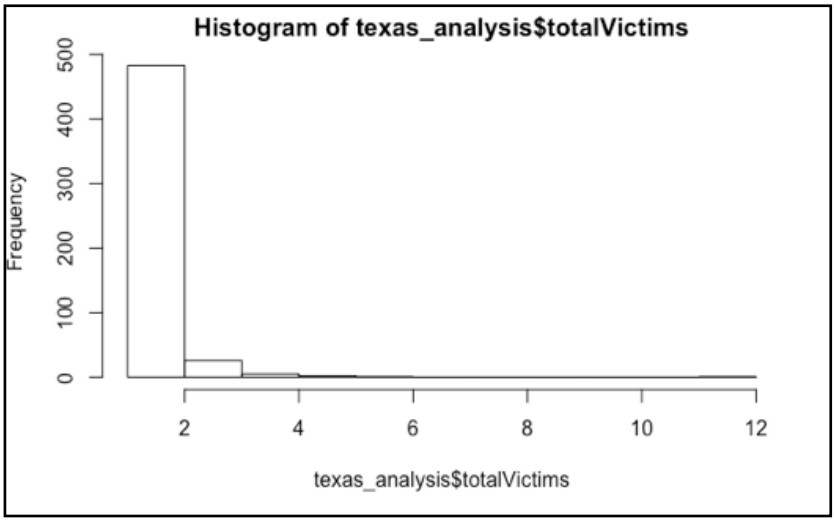


Appendix 3: EDA Histogram of 'educationYears' Variable

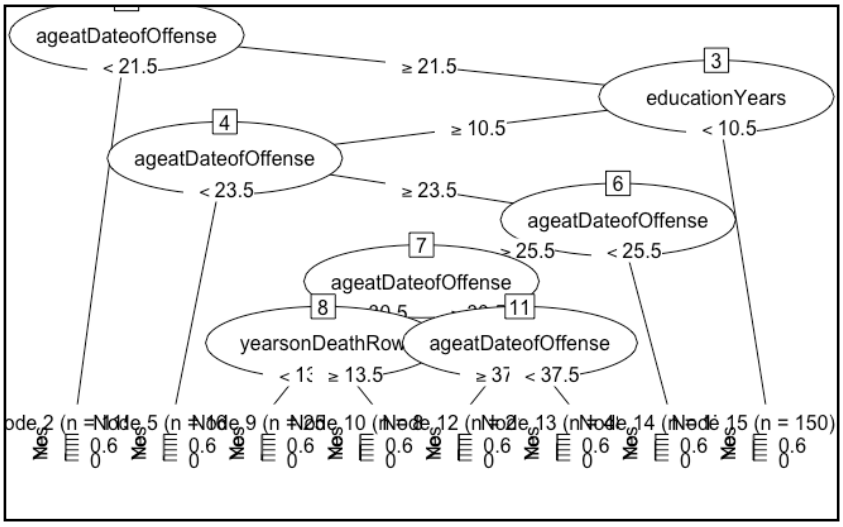


Gina O'Riordan
Behavioral Data Science - East
Final Project 12.1.18

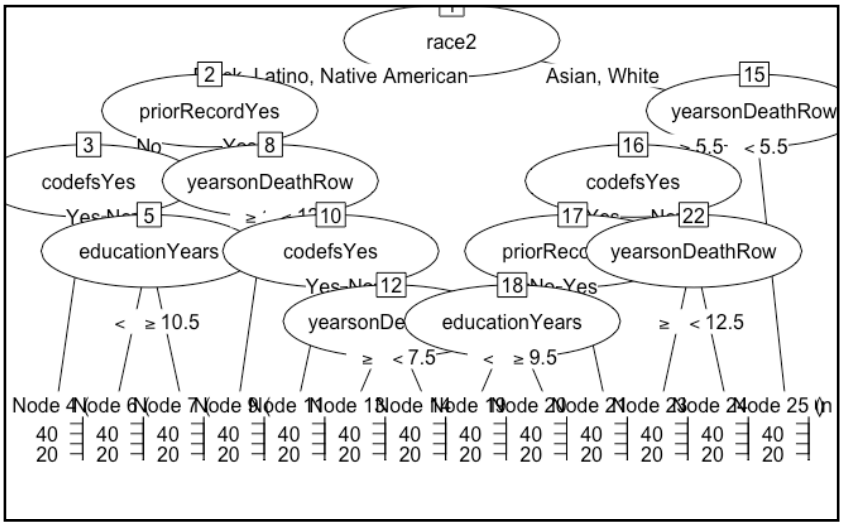
Appendix 4: EDA Histogram of 'totalVictims' Variable



Appendix 5: (tree_prior_1) Unpruned Decision Tree Predicting 'priorRecordYes' Variable



Appendix 6: (tree_age_1) Unpruned Decision Tree Predicting 'ageatDateofOffense' Variable



Gina O'Riordan
Behavioral Data Science - East
Final Project 12.1.18

Appendix 7:
(tree_years_1) Unpruned Decision Tree Predicting 'yearsonDeathRow)