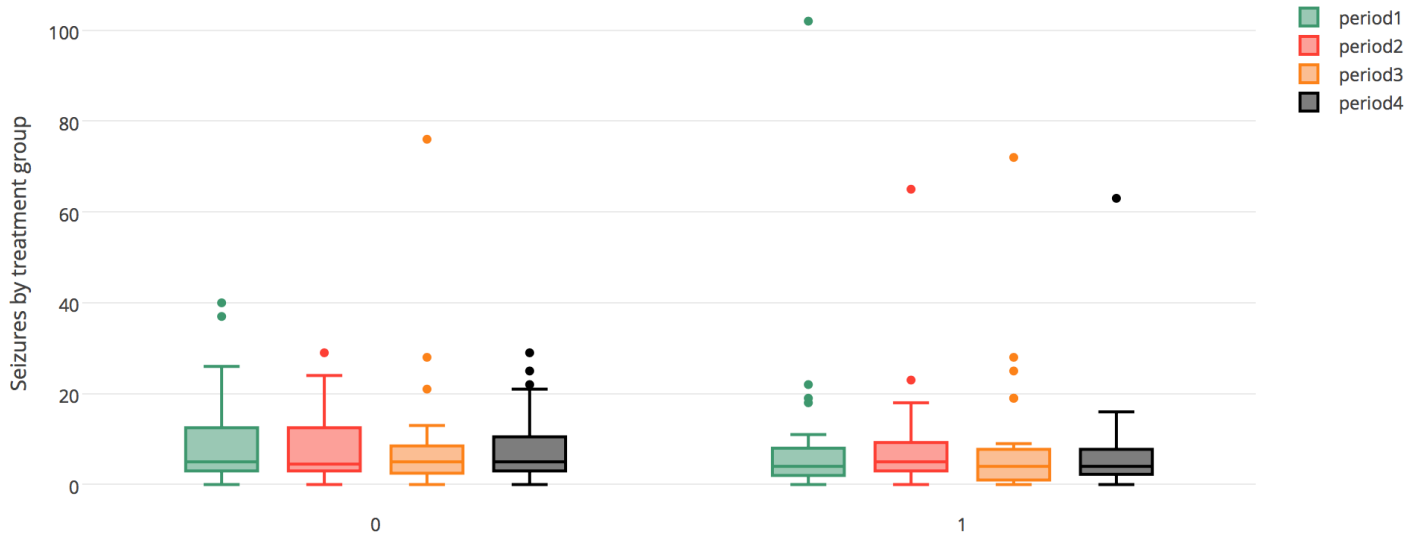
Milestone 4: Baseline Model

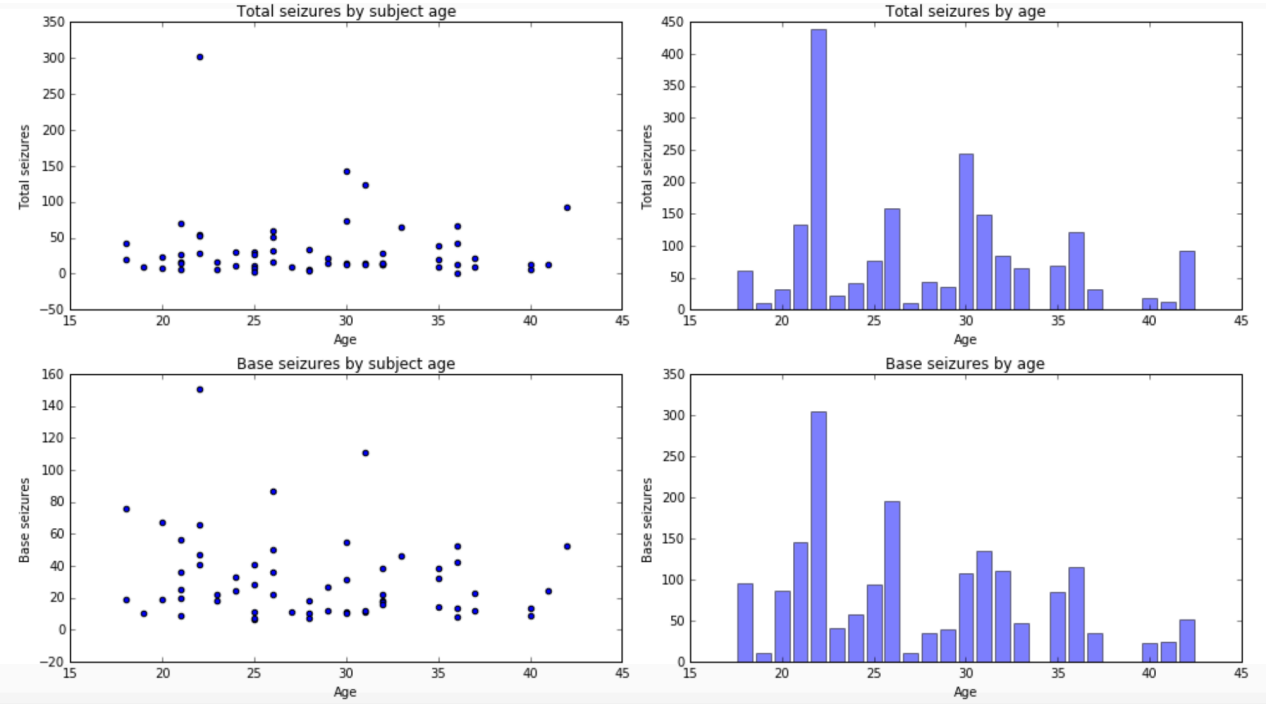
Prompt: Fit a specific model provided in your project topic's guidelines, provide the results from Python along with a roughly half page write-up of what you notice. Visuals are helpful here too.

In hoping to investigate the factors that affect epilepsy recurrences and predicting future seizure incidences, we build our baseline model using a data set with a two-week seizure counts for 59 epileptics. The number of seizures was recorded for a baseline period of 8 weeks; patients were randomly assigned to a treatment group or a control group afterwards, and counts were then recorded for four successive two-week periods.

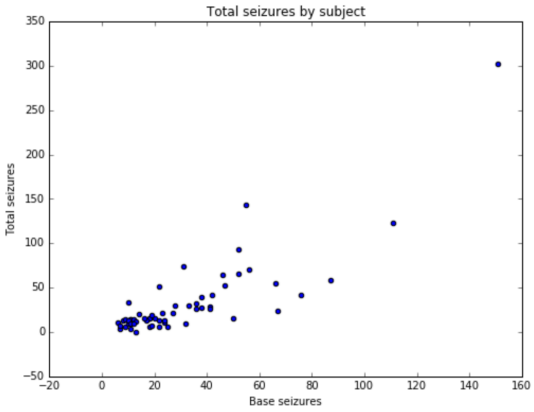
We examined the seizure incidences during the follow-up four successive periods, and find that each follow-up period and each treatment group had similar median number of seizure:



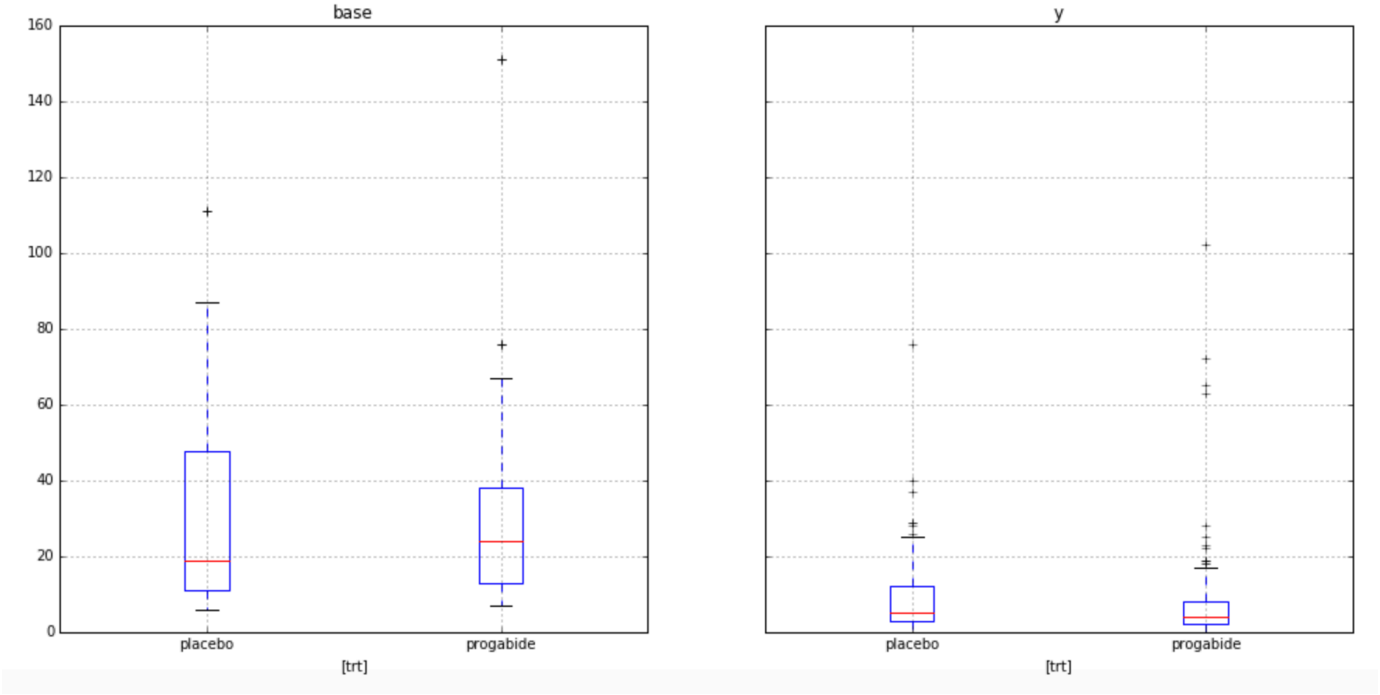
We failed to identify a seizure distribution pattern with respect to patient's age:



However, it seems that baseline total number of seizures is related to the number of seizures happened at the follow-up period.



One can possibly argue that the treatment is useful because it seems that the median number of seizure incidences is slightly higher among patients in the progabide group at baseline, but after certain treatment, the number of seizure between the two groups tend to be the same.



Base on our findings, we performed a series of variable selection methods that we learned in class, and considered these baseline models: linear regression, quadratic regression, lasso, and ridge with different tune parameters, λ. After splitting data into training and testing sets, we find that although the accuracy of quadratic regression on training set is high, its accuracy on testing set is pretty low, which indicate over fitting. Therefore, linear regression is a better model, however, the accuracy for train set is only 0.69, while that for test set is only 0.27. The unpromising accuracies are possibly due to small sample size, shortage of variables, non-statistically significant covariates, etc. Nevertheless, this baseline model can be used as a sanity check, later on, we will clean up the NCDS dataset, and run more sophisticated models on that dataset, we are optimistic with developing much more reasonable models in the future.