Collin Giguere

After using PowerTransform to see what transformations, if any, we should make to our predictor variables, we used PowerTransform again to see if we should transform our response variable. R suggested we use $Log(Crime)$ instead of $Crime$, so we tried it and it helped out the model quite a bit. We then went back and looked at the newly transformed variable NW. We didn't use it in our previous model because it didn't really have any relationship with $Crime$, but since we transformed them both we decided to go back and have another look. We looked at the added variable plots for $log(Crime) \sim Predictors$ and $log(NW) \sim Predictors$ and saw a decent relationship, so we added $log(NW)$ to our model. After looking at our final model, we realized that *So* was not significant to any degree and quickly realized it was due to our audition of $log(NW)$. Looking at the plot of $Crime \sim NW$ with colors indicating *So*, it was quite obvious that if we know that $NW < 12$ then the state is northern. This means that knowing $log(NW)$ will tell us the value of *So* in all but three cases. As a result we dropped *So* as a predictor. Our final plot of $log(Crime) \sim Fitted\ \ Values$ shows a much better linear correlation, but there was a small problem; the graph quite clearly had a sinusoidal or polynomial pattern. This is also quite clear in the graph of $Residuals \sim Fitted$. We tried sinusoidal regressions, splines, polynomial regressions, and ARIMA, but nothing we tried could help solve our problem. Lastly, we colored out data points by *So* and saw the same polynomial pattern in both the southern and non-southern states, so we knew we made the right choice in dropping it as a predictor. We think that we could have done better finding a model to predict the crime rate of a state if the data had not been compressed into less specific variables, or if we could do more research into what was causing polynomial pattern in our model.