- Looked at variables **Cause_Desc, Type_Desc, Cause_Type**
- Determine if we should use R tm text miner
- As most of our text fields are structured data, it would not make much sense to use R tm miner which is more for free text analysis,JMP text analysis and partitioning tree is more than enough
- Used partition tree with variables **Cause_Desc, Loss_Desc, VehAge_Current, AF,Veh_Make** , Response variable **Claim_Paid**
- Observed from the Actual vs Predicted plot that many of the actual claim amounts are more than the predicted amounts.
- R square or best split is 40 splits. R square very incrementally increases
- Tried to Log the **ClaimPaid** amount, following literature review to normalize the distribution
- Analyzed the Normal and Log normal goodness of fit for transformed **ClaimPaid,** both goodness of fit test have very small p-values, rejecting H0 (distribution is normally distributed)
- Tried Neural Networks, R square is 0.38, still not as good as partition tree.
- Tried to include **Repairer_Status** to see if R square would improve, did not change much
- To consult Prof regarding results of prediction
- To continue finishing up remaining exploratory data analysis