Elias Van Sickle and Alfred Hurley (Team B)

Team B ultimately decided that the best way to go about creating this model was a modified Ridge regression. Given that our outcome variable, count, was necessarily valued greater than or equal to zero for all rows, and that it essentially represented an arrival process, we changed the model to the Poisson family. By doing so, we were using a method that most closely applied to the data at hand and removing the possibility that our model would create negative predictions. Additionally, we incorporated principle component analysis to slightly simplify the model.

We first started looking at Ridge to try and gauge the importance of different variables. After looking at a few other models and deliberating about their applicability we realized Ridge would end up being the best. During our exploratory data analysis we saw that every variable seemed to have some value, so it would behoove us to use all of them. During our EDA we also noticed that atemp ('feels like' temperature) and temp (actual temperature) were highly correlated. That prompted us to do a principle component analysis and test the model using the single combined variable against the model with both variables. We were happy to see a slightly improved RMSLE with the simpler model and therefore incorporated the PCA.

An unforeseen issue in the submission of the model was the formatting of the dates. Kaggle required a very specific format that is unavailable in any package that we could find online. This hurdle forced us to be creative. Our solution incorporated material found online as well as our own hack. Broadly, this problem was representative of one our learning processes: a large part of self-sufficiency as a coder is being able to sift through the resources of the Internet effectively and use the most applicable insights as aid.