Chapter 4 Homework

Info 3130

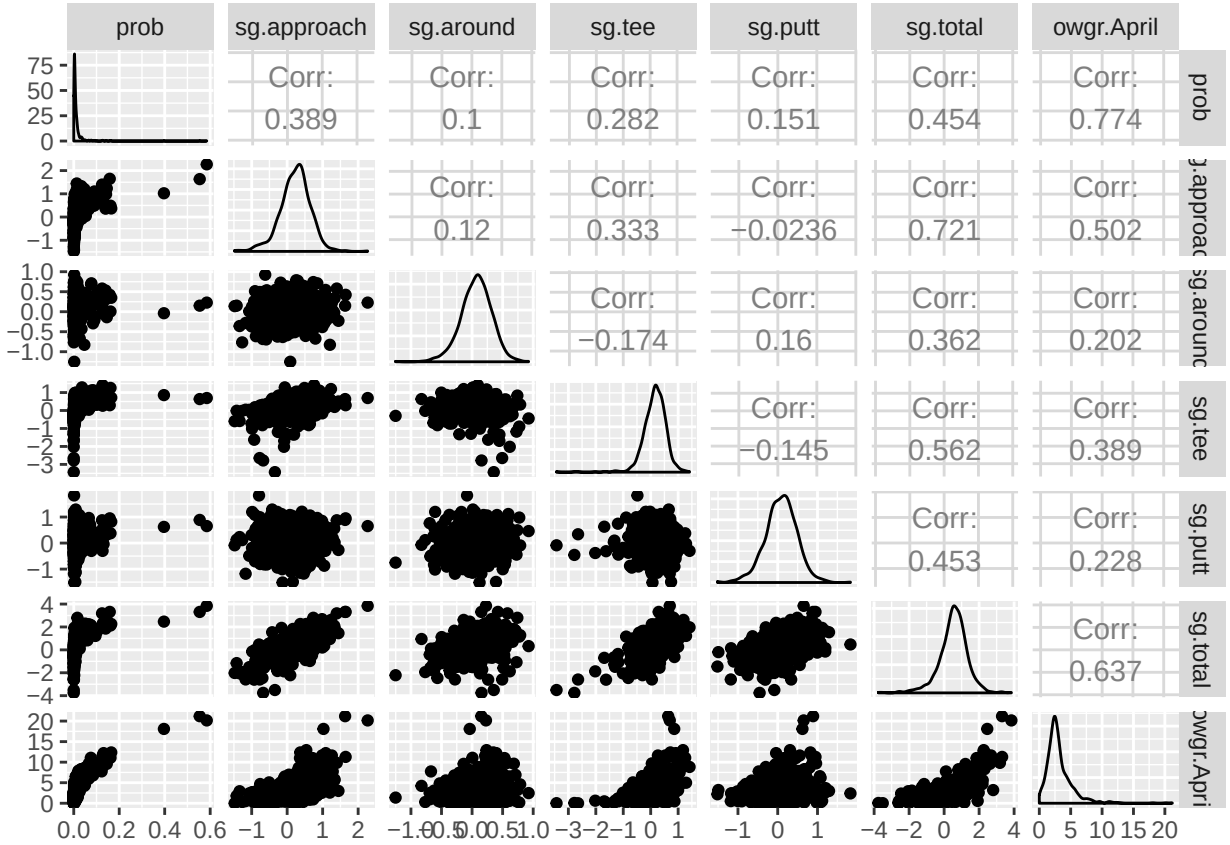
Cody Frisby

9/18/2017

PGA Tour Data

My data comes mainly from the PGA Tour. Included in my data set are also the results of simulations run using probabilities generated from fitting an proportional odds model to a set of predictor variables. Also included are the Vegas Odds for the Masters golf tournament for each year, if available, the *actual finish* for each player each year, and their official world golf ranking in April and February of each year. Although the years included in the data set are 2004 - 2017, the Vegas odds were only able to be gathered going back to 2014. The **sg** catagories are calculated by the PGA tour for each player who appears on the PGA tour. They are a measure of the individual players skill in the different catagories compared to his peers.

For the correlation matrix below I've included the variables *prob*, *sg.approach*, *sg.around*, *sg.tee*, *sg.putt*, *sg.total*, and *owgr.April*.



The largest correlation exists between *owgr.April* and *prob*. Recall that *prob* is the modeled/simulated probability of a given player winning the Masters. This large correlation indicates the large influence that a players Official World Golf Ranking has on the modeled/simulated probabilites.

The next largest correlation exists between *sg.total* and *sg.approach*. This makes sense since *sg.total* is the sum of all the **sg** catagories. Additionally, if a player is better than his peers in the "approach" catagory then

that would directly affect their "total" **sg** score.

There appears to be some outliers among the *prob* variables. When investigating this further, it appears these are mainly Tiger Woods for a few years that he was extremely dominant.

If we zoom in on on just one of the bivariate plots we can see how dominant Tiger Woods was for a few years. I've also added some coloring where we can see who actually won the masters each year in our data set (blue triangle).