[Data Driven March Madness] Milestone#3

Robert Shaw, Sean Coleman, Spencer Evans, Daniel Alpert

Exploration Process

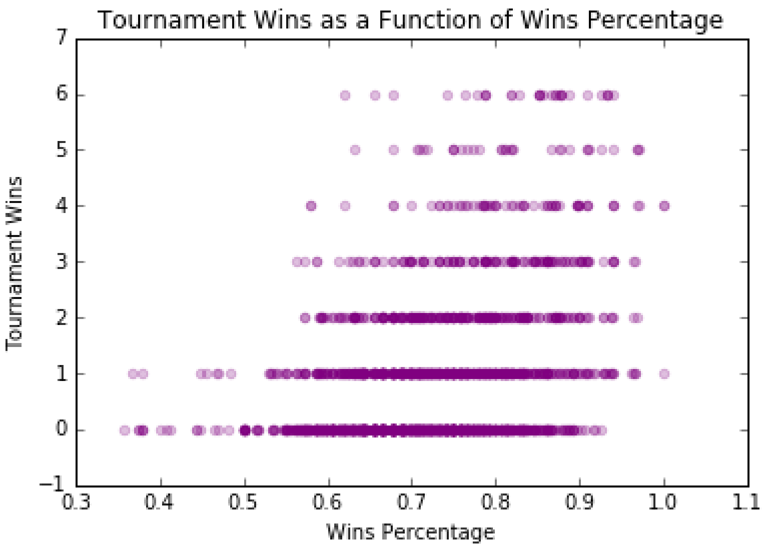
In this phase of the project, we cleaned the kaggle dataset and investigated some possible predictors to include in our model. Since the kaggle database contains a list of games played over the years, without any sort of summary statistics, we decided to create a variety of predictors to test. We created the 11 variables for all teams over the regular season data from 1985-2016. Although our actual model will eventually predict the probability of a team winning a particular game, we have not made this model, yet. However, in making these variables we wanted to test which ones are useful predictors of tournament success. As such, we used a proxy to identify predictors which may be important in our model: total tournament wins. This response is a great proxy for the probability a team will win a particular post season game, becuase it is the ultimate metric of post-season success. The variables are listed below, coupled with several graphs of the correlations between the individual predictors and number of tournament wins. However, it is important to note that these variables will be highly coorelated, so even if we see a relationship with the response, it does not mean that these variables will have predictive significence on their own. As such, we will have to use cross validation in our variable selection. These are simply candidate variables, which we will test.

[Data Driven March Madness] Milestone#3

Robert Shaw, Sean Coleman, Spencer Evans, Daniel Alpert

Prospective Variables

```
1) Regular Season Win Percentage
```

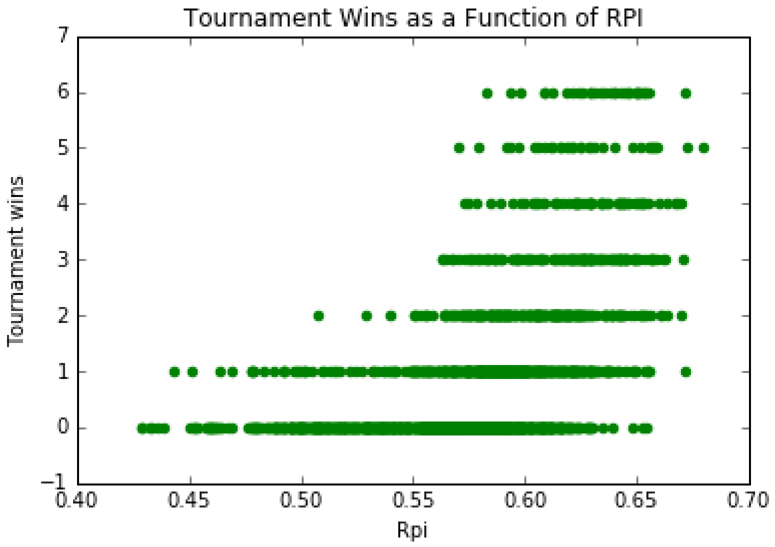


This graph plots tournament wins as a function of regular season win percentage. We see that nearly all teams in the tournament have regular season win percentages above 50%. And while some teams can make it to the final four with win percentages below 70%, none have managed to win the tournament. Conversely, we see that there are a number of teams with regular season win percentages above 95%. However, none of those teams ended up winning the ship. In fact, only one or two of them made it to the final four. This could occur because of teams in weak conferences breezing through an easy regular season schedule. It could also be indicative of a team that hasn't faced adversity.

[Data Driven March Madness] Milestone#3

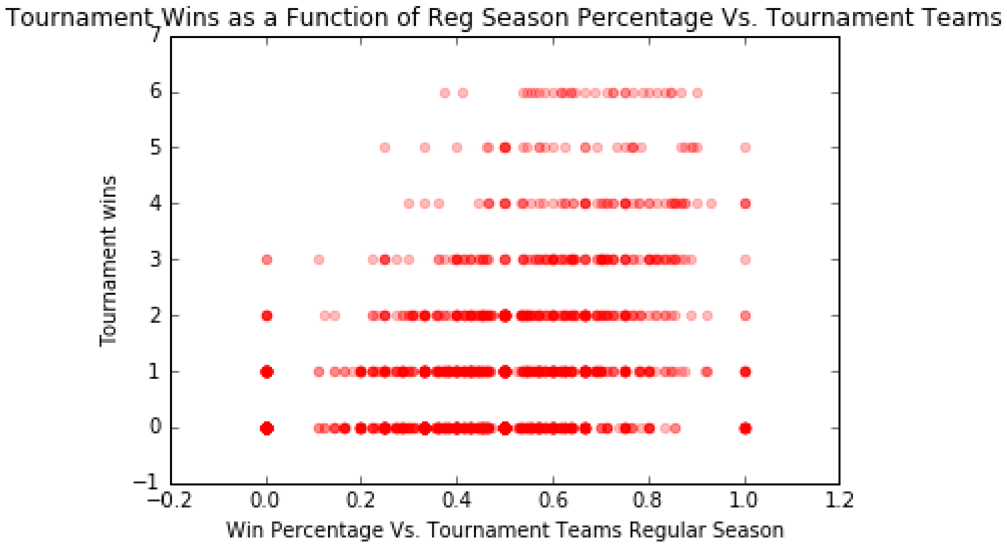
Robert Shaw, Sean Coleman, Spencer Evans, Daniel Alpert

```
2) RPI (Strength of Schedule)
```



As we can see, there appears to be some correlation between RPI and tournament wins, as we would expect. Thus it seems valid to consider including RPI or some other metric of strenght of schedule in our model to predict whether a team will win a tournament game.

```
3) Win Percentage Vs Tournament Teams
```

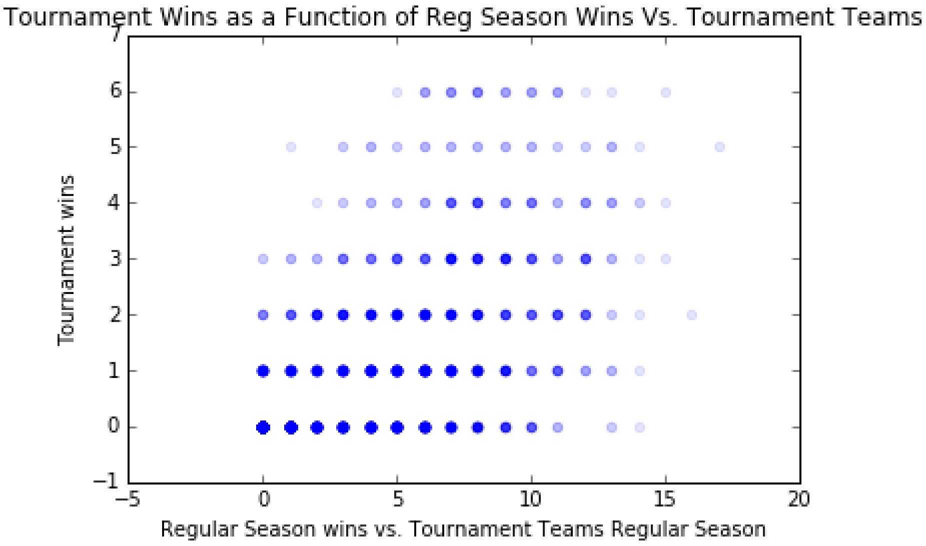


As we can see there is only a very small correlation between the win percentage in the regular season and tournament wins, which is somewhat unexpected. However, we could think that teams which win in the tournament are "battle tested" and will have played serveral regular season games against good teams, leading to a low percentage of wins versus highly ranked teams. As such, we will consider total wins instead of win percentage.

[Data Driven March Madness] Milestone#3

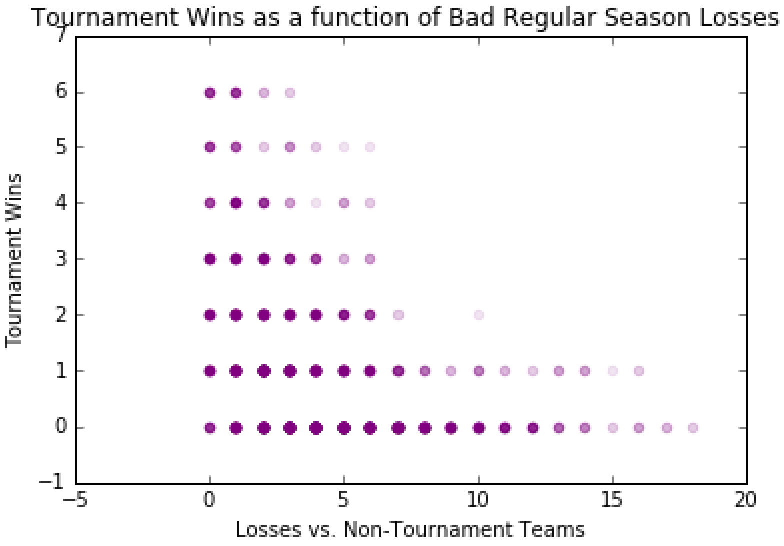
Robert Shaw, Sean Coleman, Spencer Evans, Daniel Alpert

```
4) Good Wins (Wins Vs Tournament Teams)
```



This visualization of the number of tournament wins as a function of the number of wins against tournament teams in the regular season is very hard to interpret becuase of the discreteness in the graph. However, we do see a general trend of increasing tournament wins as regular season wins vs tournament teams increases. As such, we will need to use a more emperical approach to analyzing the effect of including this predictor in our model: namely, via cross validation down the road.

```
5) Bad Losses (Losses Vs Non-Tournament Teams)
```

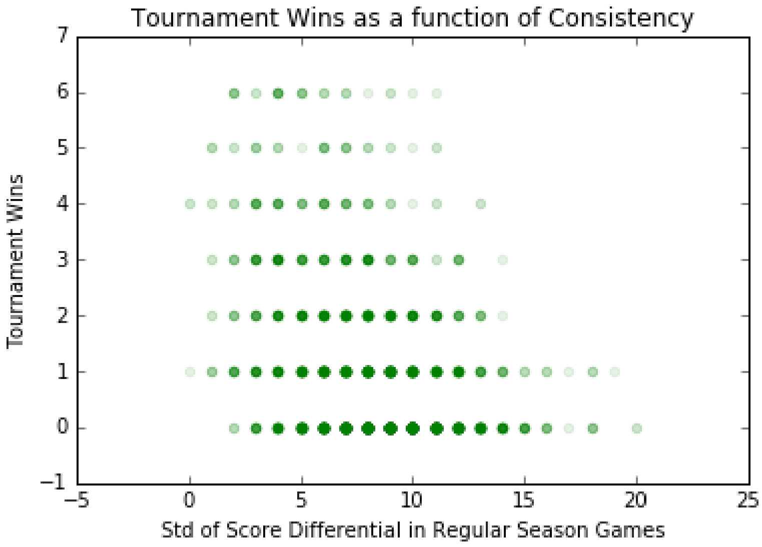


As we can see, there is a strong negative correlation between tournament wins and losses to non-tournament teams in the regular season. Thus, the predictor of "Bad Losses" should probably be in our model, subject to our empircial tests using cross validation. Teams with a low number of bad losses are more likely to win tournament games.

[Data Driven March Madness] Milestone#3

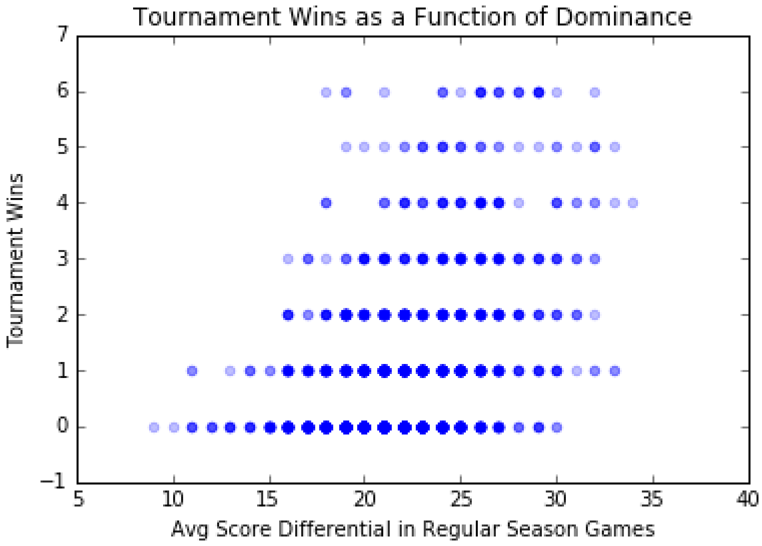
Robert Shaw, Sean Coleman, Spencer Evans, Daniel Alpert

`6) Consistency (Std. Scoring Differential)`



This visualization of the number of tournament wins as a function of the consistency (std of score differential) in the regular season is hard to interpret becuase of the discreteness in the graph. There does appear to be a trend that teams with low standard deviations tend to do better than those with higher deviations, but it is hard to tell. As such, we will need to use a more emperical approach to analyzing the effect of including this predictor in our model: namely, via cross validation down the road.

`7) Dominance (Avg. Scoring Differential)`

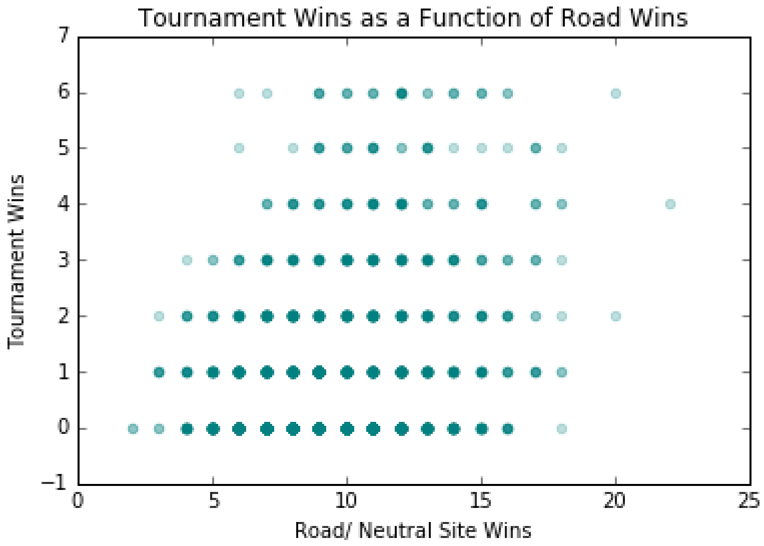


As we can see, there appears to be a slight correlation between tournament wins and a teams ability to dominate in the regular season, where domination is measured by the average score differential in the regular season. Qualitatively, we would expect this to happen, so it makes sense for us to try to include this in our model. Obviously, the inclusion of this in the model is subject to cross validation, as there may be confounding variables that are not present.

[Data Driven March Madness] Milestone#3

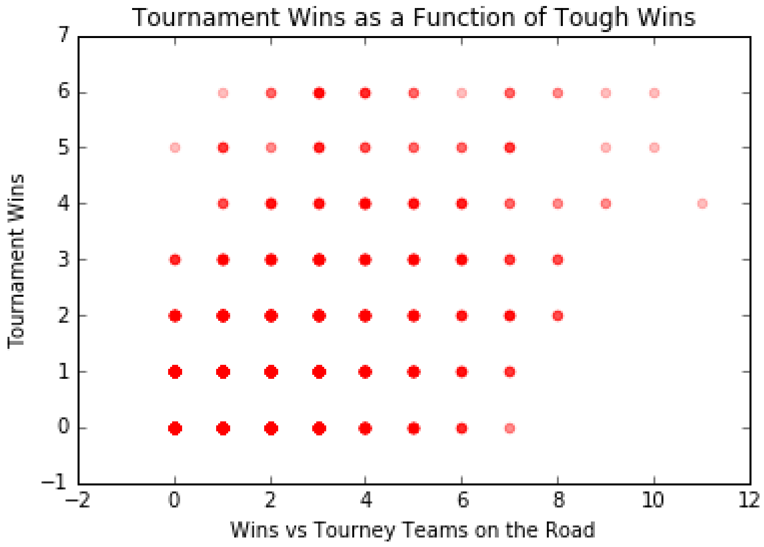
Robert Shaw, Sean Coleman, Spencer Evans, Daniel Alpert

```
8) Away Wins
```



As we can see, there is a very very slight coorelation between away wins and tournament wins. However, it is somewhat hard to visualize the relationship becuase of the discretness, so we should consider empirical approaches such as CV before rejecting Road Wins from the Model. Further, we should also consider "tough wins," which are road wins against tournament teams.

```
9) Tough Wins (Away Win Vs Tournament Teams)
```

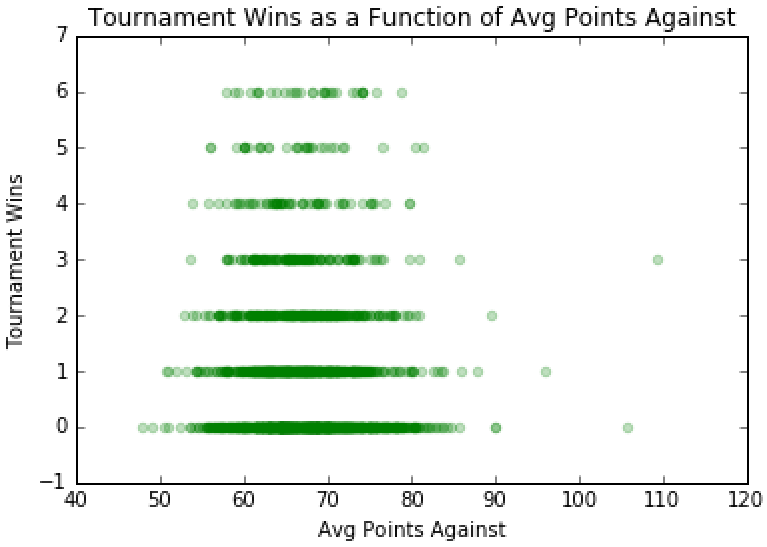


Once Again, due to the discreteness in the graph, it is hard to visualize the relationship between tournament wins and tough regular season wins (road wins vs tournament teams). As such, we would consider using an empircial approach (such as CV) to understand the importance of this predictor. Since we qualitiatively understand that this metric should be a predictor, since winning games against top teams on the road should relate to tournament wins. Thus, we will put this predictor under further review.

[Data Driven March Madness] Milestone#3

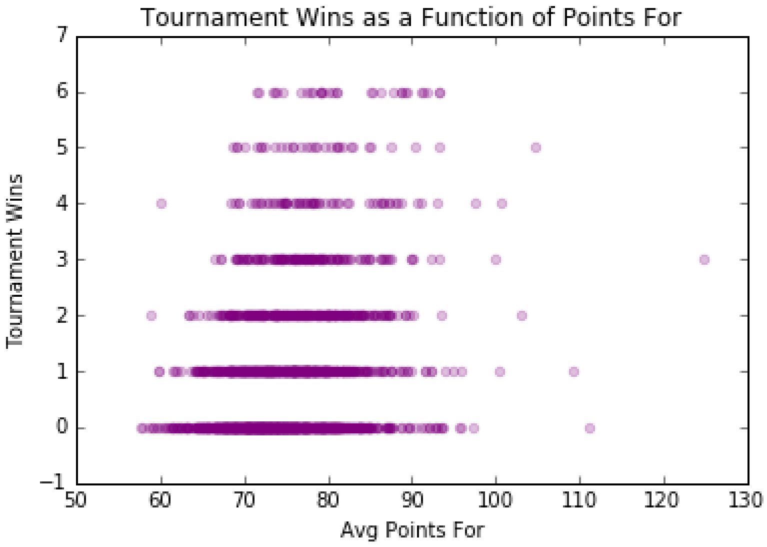
Robert Shaw, Sean Coleman, Spencer Evans, Daniel Alpert

```
10) Defensive Prowess (Avg Points Against)
```



There does not seem to be any soorelation between average points against and tournament wins. This makes sense, becuase defensive teams and offensive teams can both be successful in the tournament. However, we should still test this predictor with cross validation.

```
11) Offensive Prowess (Avg Points For)
```



There does not seem to be any coorelation between average points for and tournament wins. This makes sense, becuase defensive teams and offensive teams can both be successful in the tournament. However, we should still test this predictor with cross validation.

Moving forward

Finally, although these are the variables we have identified thus far, we are also considering using some metrics outside of the kaggle database about the coaches and players of each of the teams. Specifically, we will be looking at the experience of the coaches and impactful returning players in the tournament to see if there is some relationship. We will also be looking at the momentum of teams who are peaking at the right time in the season. We will also be looking at teams with star players, capable of taking over a game. These are just some initial ideas to begin with, we will be exploring many more variables as the process continues.