Karl Shiffler – Project on March Madness NCAA Tournament Data

I retrieved my data from the Kaggle March Machine Learning Mania competition (https://www.kaggle.com/c/march-machine-learning-mania-2015). The data took the form of multiple csv files which I needed to reorganize. The data available covered the seasons and tournaments for the years 2003-2014, as well as the 2015 regular season data. For regular season games and pre-2015 tournament games, we have access to per-game statistics: field goals made, field goals attempted, three pointers made, three pointers attempted, free throws made, free throws attempted, offensive rebounds, defensive rebounds, assists, turnovers, steals, blocks, and personal fouls for each game. Most of these statistics generally follow a normal distribution and there is no missing data. We also have access to tournament rankings. I have summarized the per-game tournament statistics data below by reporting mean and standard deviations.

FG = Field Goal
FT = Free Throw
O = Offensive
D = Defensive
TO = Turnover

For winning teams:

	score	FG made	FG attempt	3pt made	3pt attempt	FT made	FT attempt
Mean	74.6102	25.9858	54.8423	6.7538	17.4653	15.8846	21.8846
Std Dev	10.6864	4.7766	7.1855	2.8135	5.2325	6.1456	7.7860

	O Rebounds	D Rebounds	Assists	TOs	Steals	Blocks	Fouls
Mean	10.9705	25.7397	14.3782	11.9884	6.5525	3.9128	16.4730
Std Dev	4.0205	5.0018	4.2528	3.8588	3.0303	2.6238	3.8647

For losing teams:

	Score	FG Made	FG Attempt	3pt Made	3pt Attempt	FT Made
Mean	63.1884	22.6961	57.4692	6.0012	19.7705	11.7948
Std Dev	10.4897	4.1776	7.6786	2.7411	5.7481	5.1202

	FT Attempt	O Rebounds	D Rebounds	Assists	TOs	Steals	Blocks	Fouls
Mean	17.0	11.4794	20.9820	11.4961	12.4717	5.9397	2.9179	19.1230
Std Dev	6.6245	4.2069	4.1463	3.6565	3.9675	2.7573	2.0627	4.2447

Possible Hypotheses and how I intend to test them:

Which stats most influence winning (besides score)?
Compute regressions on the game data with the target being a win or loss.

Is regular season performance a reliable indicator of tournament performance? How much predictive power does it have?
Compare the regular season and tournament statistics for the same teams in the same year. Perhaps train a classifier on a team's regular season data and then test it on their tournament matchups.

How do different stats change over the years? Why? How do different stats' predictive power change over the years?

Look at averages and how they change over the years. Perhaps research rule changes or new players that have changed the way the game was played to explain these changes. We can then investigate how these paradigm shifts affect the predictive power of these statistcs.

High volume of shot attempts vs efficiency? What's most effective and has that changed over the years? Compare predictive power in regressions of accuracy percentages (FG made/FG attempted) to number of shots taken. We can also track this through different seasons.

Conventional basketball wisdom claims that free throw percentages make or break teams. To what extent is this true, and has it changed over time?
Similarly, rebounds (missing defensive and allowing offensive) is often purported to be the deciding factor in games. Look at predictive power and disparity between winning and losing teams on average. Also look over time.
We can look at the coefficient of these numbers in a regression and also could attempt to calculate a regression based solely on free throw numbers (or rebound numbers) and see how well it can split the game results.

How many fouls are committed/called during regular seasons vs tournaments? Over time? How large is officials' effect on play?
This will be especially interesting to look at in the context of historical rule changes.

Better to have a few stars or many team players?
Look at assists vs unassisted FGs (FG made - assists). Again, we can look at the coefficients of these values in a regression or try to train a classifier based strictly on these numbers and see to what degree it can predict the outcome.