Wrangling: You have three files, Teams.csv, RegularSeasonDetailedResults.csv and TourneyCompactResults.csv. The first gives a list of the team numbers (the names are not important), the second gives the results of each game of the season for each team, and the third gives a (compact) record of the result of each tournament game. Your first goal is to create a file that contains, for each team and each season, a record of the team's *statistics* for that season. Here *statistics* refers to the average value for each of the stats contained in the data dictionary. So a row in this dataset should look like this:

Season	Team	Avfgm	Avefga	Avefgm3 ….	Avepf
2003	1101	23.4	48.4	10.2 …	6.3

To create this, you will need to use the Team number (found in the Teams.csv file) and search the RegularSeasonDetailedResults.csv file for all games played by that team for each season 2003 – 2017. Keep in mind the team can appear as either the winning team (Wteam) or losing team (Lteam). Once you have all of the games the team played, average the columns for the statistics to get the entries in the data file you are creating. Loop over seasons 2003-2017 and all teams.

Clustering: Perform a cluster analysis on the file you created, and see if you can cluster teams. Remove the Season and Team columns before you do this.

Regression: You now want to build a model that will try to predict the number of points the team will score in the tournament based upon their season average. In the TourneyCompactResults.csv file is record of every tournament game since 1985; you only have season averages starting in 2003, so you will only be able to use the games from 2003 on. For each of the games in the TourneyCompactResults.csv file from 2003 on, you have a winning team, winning team score, losing team, and losing team score. Hence, for each game in this file, you will get **two** rows in your new dataset (one for each team). So for example, the first two rows will probably look something like this:

Season	Team	Avfgm	Avefga	Avefgm3 ….	Avepf	Points
2003	1421	23.4	48.4	10.2 …	6.3	92
2003	1411	26.3	46.8	16.3 …	8.3	84

….

You will get the Season and Team numbers from TourneyCompactResults.csv, and the average numbers from the dataset you created in the wrangling. Using this new dataset, you should be able to build a predictive model – again, drop Season and Team, and just use the average statistics. The *Points* value is your response variable, this is the number you want to predict. So basically, given a team's average stats for the year, your model should predict how many points they will score in the tournament game.