DATA_ADJUST

November 17, 2017

Darren Lund

In [3]:
```
import numpy as np
import pandas as pd
import os
from matplotlib import pyplot as plt
```

In [39]:
```
# Load DATA
path = './DATA/2017/Gonzaga'

# Split function (for 2-pt, 3-pt, and ft)
def split(ratio) :
    values = [value.split('-') for value in ratio]
    made , attempt = [int(shots[0]) for shots in values], [int(shots[1]) f
    return made, attempt

# Different types of shots
shot_types = ['2Pt','3Pt','FT']

# Walk through player files
for dir_path , dir_name , file_names in os.walk(path) :
    # List of players
    players = {}
    for name in file_names :
        # Only worry about cleaned data
        if name[-3:] == 'csv' :
            # Don't get the Scouting report
            if name[:8] != 'Scouting' :
                players[name[:-6]] = pd.read_csv(os.path.join(dir_path,nam
    team_values = {}
    # Add %Att for 2s, 3s, and FT
    for player in players.keys() :
        for shot_type in shot_types :
            made , attempt = split(players[player][shot_type].values)
            if shot_type in team_values.keys() :
                team_values[shot_type] = [team_values[shot_type][i] + atte
            else :
                team_values[shot_type] = attempt
```

```
for player in players.keys() :
    for shot_type in shot_types :
        made , attempt = split(players[player][shot_type].values)
        perc_att = [attempt[i] / team_values[shot_type][i] for i in ra
        players[player][shot_type+' %Att'] = perc_att
    # Add apprx points prevented from blocks and steals
    points_prev = [2*(players[player].loc[i]['Blk'] + players[player].
    players[player]['Pnts-Prev'] = points_prev
    # Add point margin
    res = players[player]['Result']
    result = [res[i][0] for i in range(len(res))]
    res = [res[i][3:] for i in range(len(res))]
    score_1 , score_2 = split(res)
    margin = [abs(score_1[i]-score_2[i]) if result[i]=='W' else -abs(s
    players[player]['Marg'] = margin
    players[player].to_csv(os.path.join(dir_path,player)+'_adj')
```

The columns I added to individual player data are the percent of the team shots each player made (for 2s, 3s, and free throws), an approximate points prevented variable (calculated by multiplying number of blocks and number of steals by 2), and a margin variable indicating how many more (less if negative) points the team had at the end of the game. The reason for the first three is so that I can look at approximately how many shots each individual player contributed per game of all kinds, so to better understand their offensive contribution to the game as a whole. In like manner, points prevented is to better understand the defensive contribution each player made. The margin variable makes it easier to compare the overall result of the game (win or lose by x points) together with each player's individual contributions.

I had to redo a bit of my data cleaning because I realized that I dropped games where a player didn't play at all, which threw off my numbers a bit. To remedy this, I went back and switch all games that were "Did not play" with a stat of either 0 or $0 - 0$. While I'm currently not sure how exactly I want to handle games that they didn't participate in, I feel this is the best way to currently store the information. It's simple enough to fix if later on I decide that it needs to be a different value.

In [ ]: