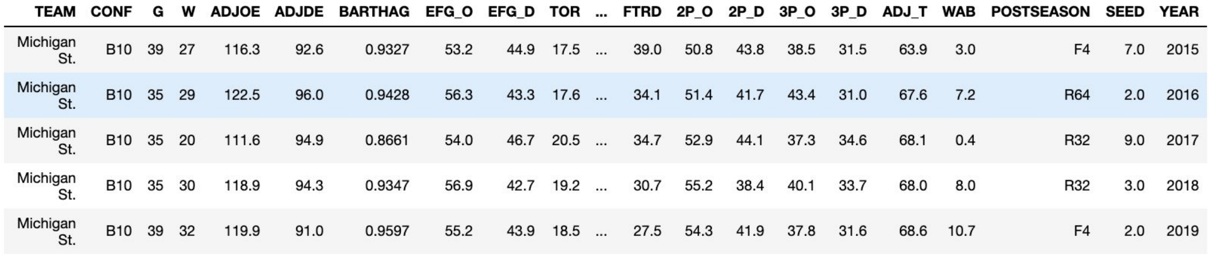
CMSE 202 Final Project Sports 2
Group 15

Nick Mouaikel, Zane Shango, Tony Tedesco

Question

- Is it possible to predict a college basketball team's number of wins on a season, based on their statistics?



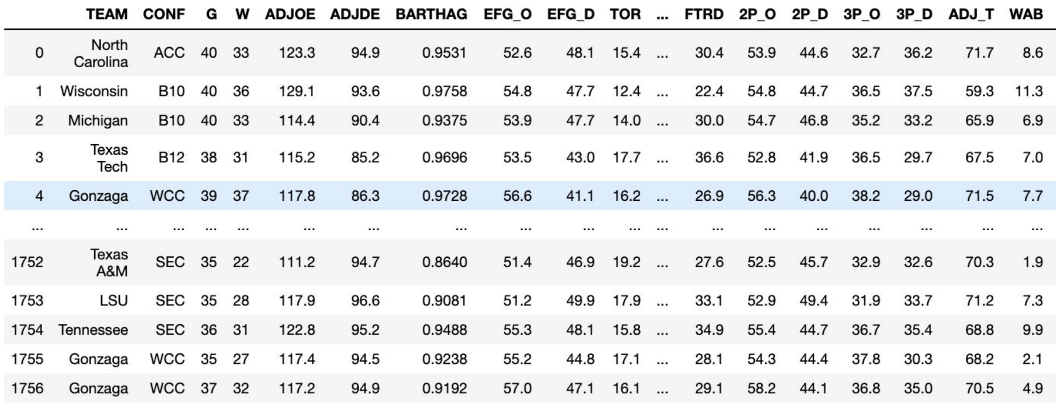
TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	...	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB	POSTSEASON	SEED	YEAR
Michigan St.	B10	39	27	116.3	92.6	0.9327	53.2	44.9	17.5	...	39.0	50.8	43.8	38.5	31.5	63.9	3.0	F4	7.0	2015
Michigan St.	B10	35	29	122.5	96.0	0.9428	56.3	43.3	17.6	...	34.1	51.4	41.7	43.4	31.0	67.6	7.2	R64	2.0	2016
Michigan St.	B10	35	20	111.6	94.9	0.8661	54.0	46.7	20.5	...	34.7	52.9	44.1	37.3	34.6	68.1	0.4	R32	9.0	2017
Michigan St.	B10	35	30	118.9	94.3	0.9347	56.9	42.7	19.2	...	30.7	55.2	38.4	40.1	33.7	68.0	8.0	R32	3.0	2018
Michigan St.	B10	39	32	119.9	91.0	0.9597	55.2	43.9	18.5	...	27.5	54.3	41.9	37.8	31.6	68.6	10.7	F4	2.0	2019

Goal

- To use machine learning to predict the total amount of wins for each team and conference in men's division 1 college basketball based on team statistics gathered over the last 6 years.

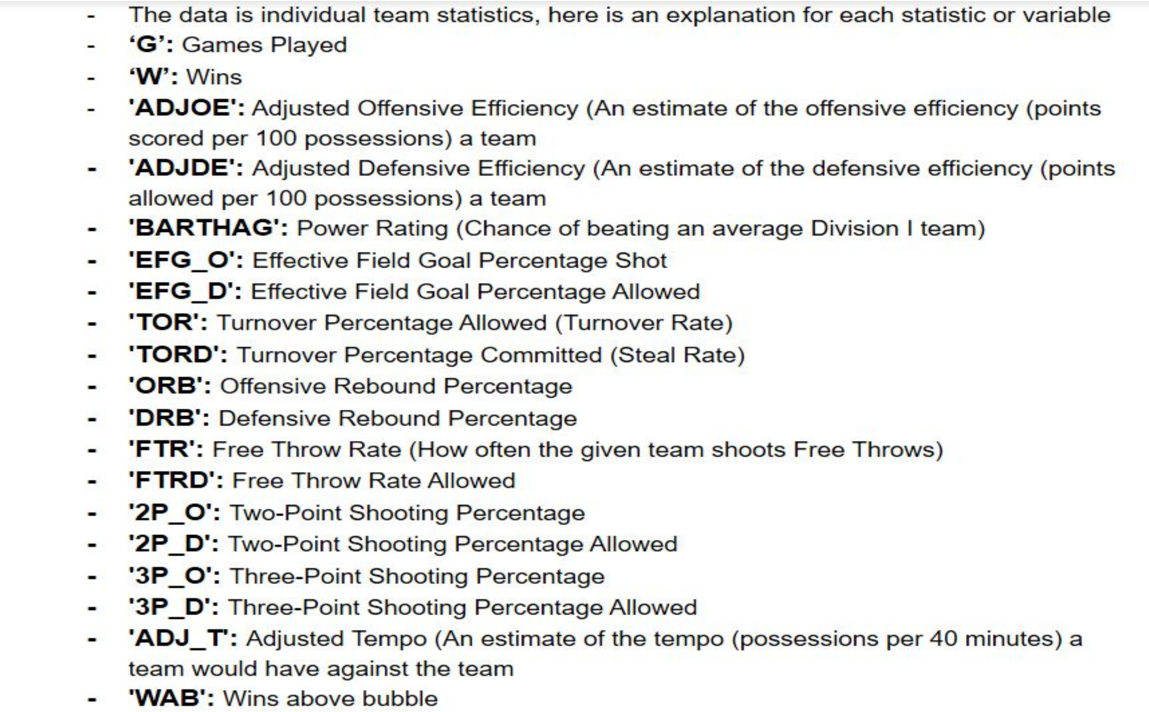
Data

- Our data is a spreadsheet that contains team statistics of all division 1 men's basketball teams over the past six years
- Data comes from the free data site kaggle.com



	TEAM	CONF	G	W	ADJOE	ADJDE	BARTHAG	EFG_O	EFG_D	TOR	...	FTRD	2P_O	2P_D	3P_O	3P_D	ADJ_T	WAB
0	North Carolina	ACC	40	33	123.3	94.9	0.9531	52.6	48.1	15.4	...	30.4	53.9	44.6	32.7	36.2	71.7	8.6
1	Wisconsin	B10	40	36	129.1	93.6	0.9758	54.8	47.7	12.4	...	22.4	54.8	44.7	36.5	37.5	59.3	11.3
2	Michigan	B10	40	33	114.4	90.4	0.9375	53.9	47.7	14.0	...	30.0	54.7	46.8	35.2	33.2	65.9	6.9
3	Texas Tech	B12	38	31	115.2	85.2	0.9696	53.5	43.0	17.7	...	36.6	52.8	41.9	36.5	29.7	67.5	7.0
4	Gonzaga	WCC	39	37	117.8	86.3	0.9728	56.6	41.1	16.2	...	26.9	56.3	40.0	38.2	29.0	71.5	7.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1752	Texas A&M	SEC	35	22	111.2	94.7	0.8640	51.4	46.9	19.2	...	27.6	52.5	45.7	32.9	32.6	70.3	1.9
1753	LSU	SEC	35	28	117.9	96.6	0.9081	51.2	49.9	17.9	...	33.1	52.9	49.4	31.9	33.7	71.2	7.3
1754	Tennessee	SEC	36	31	122.8	95.2	0.9488	55.3	48.1	15.8	...	34.9	55.4	44.7	36.7	35.4	68.8	9.9
1755	Gonzaga	WCC	35	27	117.4	94.5	0.9238	55.2	44.8	17.1	...	28.1	54.3	44.4	37.8	30.3	68.2	2.1
1756	Gonzaga	WCC	37	32	117.2	94.9	0.9192	57.0	47.1	16.1	...	29.1	58.2	44.1	36.8	35.0	70.5	4.9

Data (Variables used)



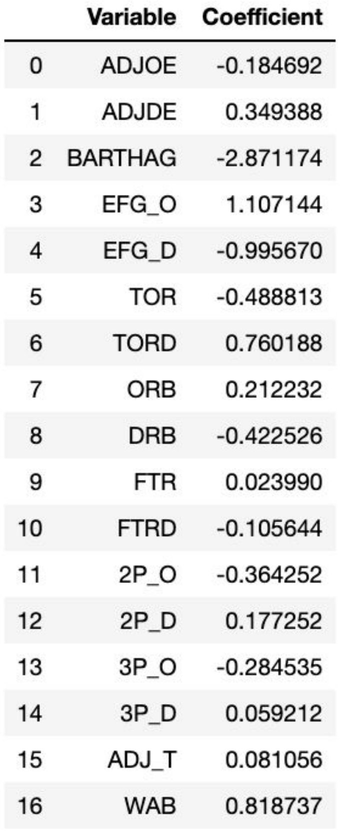
- The data is individual team statistics, here is an explanation for each statistic or variable
- **‘G’:** Games Played
- **‘W’:** Wins
- **'ADJOE':** Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team
- **'ADJDE':** Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team
- **'BARTHAG':** Power Rating (Chance of beating an average Division I team)
- **'EFG_O':** Effective Field Goal Percentage Shot
- **'EFG_D':** Effective Field Goal Percentage Allowed
- **'TOR':** Turnover Percentage Allowed (Turnover Rate)
- **'TORD':** Turnover Percentage Committed (Steal Rate)
- **'ORB':** Offensive Rebound Percentage
- **'DRB':** Defensive Rebound Percentage
- **'FTR':** Free Throw Rate (How often the given team shoots Free Throws)
- **'FTRD':** Free Throw Rate Allowed
- **'2P_O':** Two-Point Shooting Percentage
- **'2P_D':** Two-Point Shooting Percentage Allowed
- **'3P_O':** Three-Point Shooting Percentage
- **'3P_D':** Three-Point Shooting Percentage Allowed
- **'ADJ_T':** Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team
- **'WAB':** Wins above bubble

Methodology

- Using machine learning based on the aforementioned variables use the data from 2015-2019 as our train data to predict the wins for each team in 2020 and compare our accuracy to the actual total number of wins

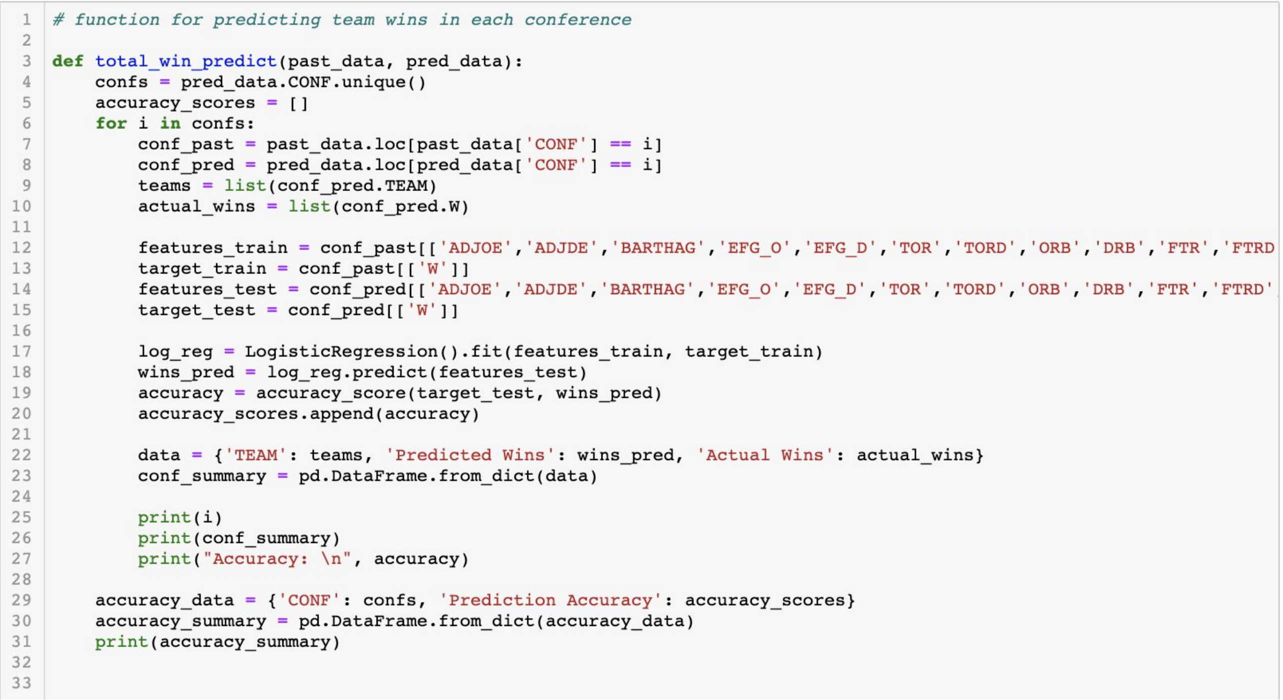
Model

- We conducted a linear regression of wins on the team stats to see which stats are most associated with wins



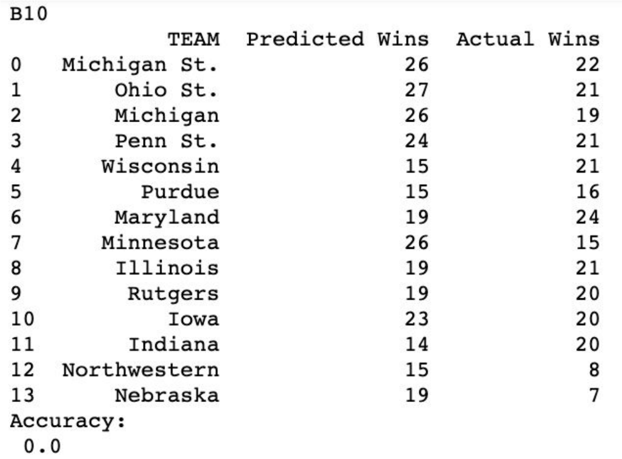
	Variable	Coefficient
0	ADJOE	-0.184692
1	ADJDE	0.349388
2	BARTHAG	-2.871174
3	EFG_O	1.107144
4	EFG_D	-0.995670
5	TOR	-0.488813
6	TORD	0.760188
7	ORB	0.212232
8	DRB	-0.422526
9	FTR	0.023990
10	FTRD	-0.105644
11	2P_O	-0.364252
12	2P_D	0.177252
13	3P_O	-0.284535
14	3P_D	0.059212
15	ADJ_T	0.081056
16	WAB	0.818737

Model

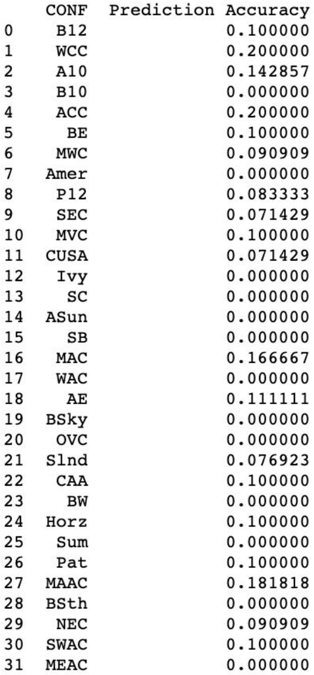


```python
# function for predicting team wins in each conference

def total_win_predict(past_data, pred_data):
    confs = pred_data.CONF.unique()
    accuracy_scores = []
    for i in confs:
        conf_past = past_data.loc[past_data['CONF'] == i]
        conf_pred = pred_data.loc[pred_data['CONF'] == i]
        teams = list(conf_pred.TEAM)
        actual_wins = list(conf_pred.W)

        features_train = conf_past[['ADJOE','ADJDE','BARTHAG','EFG_O','EFG_D','TOR','TORD','ORB','DRB','FTR','FTRD
        target_train = conf_past[['W']]
        features_test = conf_pred[['ADJOE','ADJDE','BARTHAG','EFG_O','EFG_D','TOR','TORD','ORB','DRB','FTR','FTRD'
        target_test = conf_pred[['W']]

        log_reg = LogisticRegression().fit(features_train, target_train)
        wins_pred = log_reg.predict(features_test)
        accuracy = accuracy_score(target_test, wins_pred)
        accuracy_scores.append(accuracy)

        data = {'TEAM': teams, 'Predicted Wins': wins_pred, 'Actual Wins': actual_wins}
        conf_summary = pd.DataFrame.from_dict(data)

        print(i)
        print(conf_summary)
        print("Accuracy: \n", accuracy)

    accuracy_data = {'CONF': confs, 'Prediction Accuracy': accuracy_scores}
    accuracy_summary = pd.DataFrame.from_dict(accuracy_data)
    print(accuracy_summary)
```

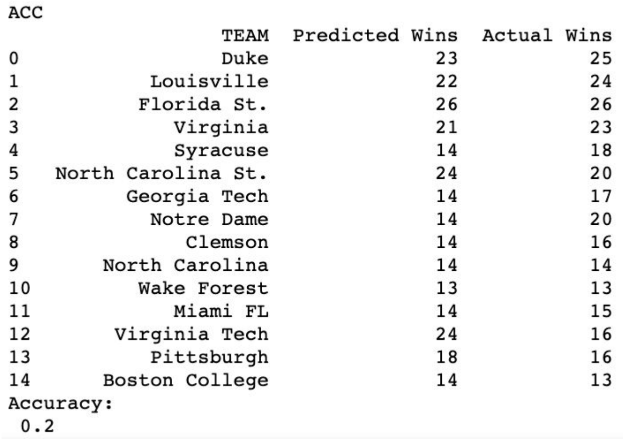
Outcome



```
B10
              TEAM  Predicted Wins  Actual Wins
0    Michigan St.              26           22
1        Ohio St.              27           21
2        Michigan              26           19
3        Penn St.              24           21
4       Wisconsin              15           21
5          Purdue              15           16
6        Maryland              19           24
7       Minnesota              26           15
8        Illinois              19           21
9         Rutgers              19           20
10           Iowa              23           20
11        Indiana              14           20
12   Northwestern              15            8
13       Nebraska              19            7
Accuracy:
 0.0
```



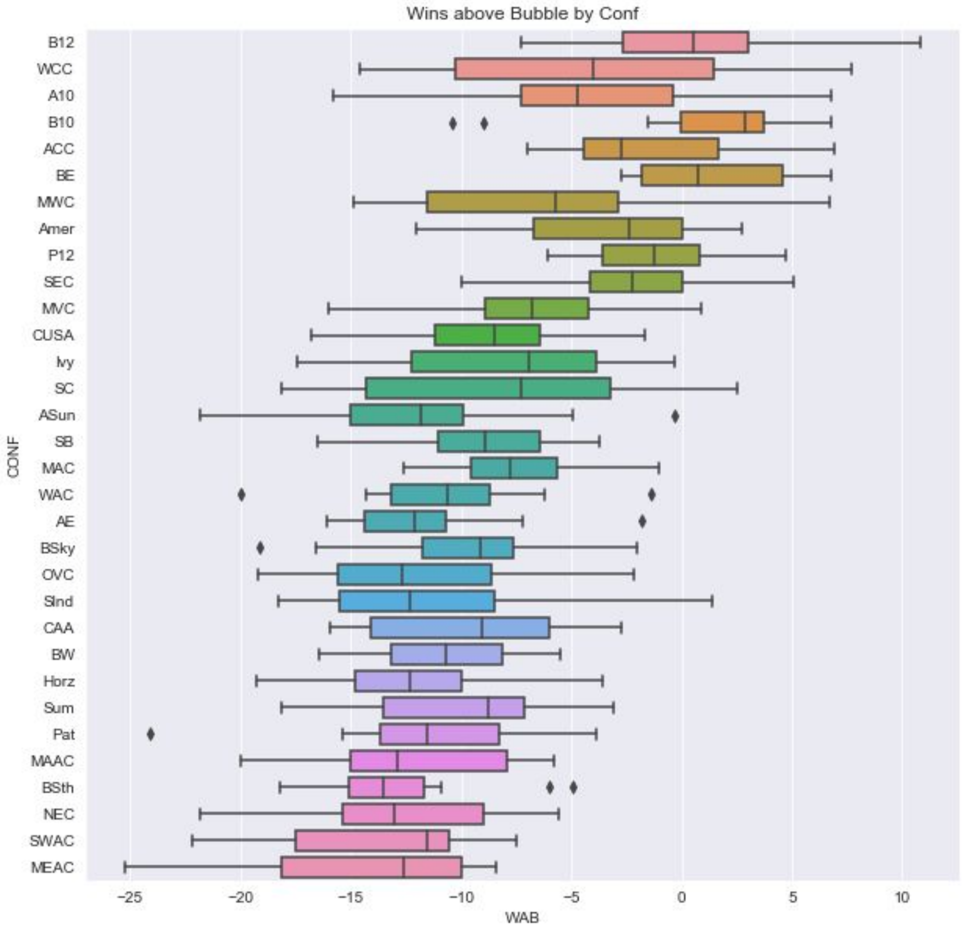
```
    CONF  Prediction Accuracy
0    B12             0.100000
1    WCC             0.200000
2    A10             0.142857
3    B10             0.000000
4    ACC             0.200000
5     BE             0.100000
6    MWC             0.090909
7   Amer             0.000000
8    P12             0.083333
9    SEC             0.071429
10   MVC             0.100000
11  CUSA             0.071429
12   Ivy             0.000000
13    SC             0.000000
14  ASun             0.000000
15    SB             0.000000
16   MAC             0.166667
17   WAC             0.000000
18    AE             0.111111
19  BSky             0.000000
20   OVC             0.000000
21  Slnd             0.076923
22   CAA             0.100000
23    BW             0.000000
24  Horz             0.100000
25   Sum             0.000000
26   Pat             0.100000
27  MAAC             0.181818
28  BSth             0.000000
29   NEC             0.090909
30  SWAC             0.100000
31  MEAC             0.000000
```



```
ACC
                  TEAM  Predicted Wins  Actual Wins
0                 Duke              23           25
1           Louisville              22           24
2          Florida St.              26           26
3             Virginia              21           23
4             Syracuse              14           18
5    North Carolina St.              24           20
6         Georgia Tech              14           17
7           Notre Dame              14           20
8              Clemson              14           16
9       North Carolina              14           14
10         Wake Forest              13           13
11            Miami FL              14           15
12        Virginia Tech             24           16
13          Pittsburgh              18           16
14      Boston College              14           13
Accuracy:
 0.2
```

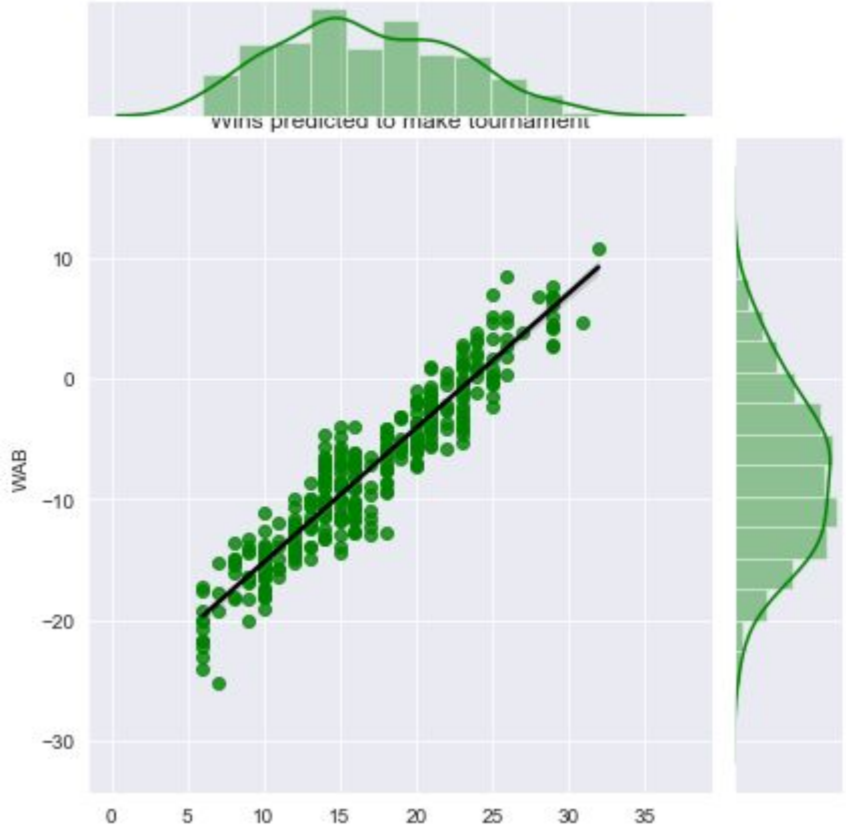
Outcome

This graph shows the average Wins above bubble (Cutoff between making the NCAA tournament and not making it). This graph shows that some conferences have a better chance at making the tournament than other conferences.



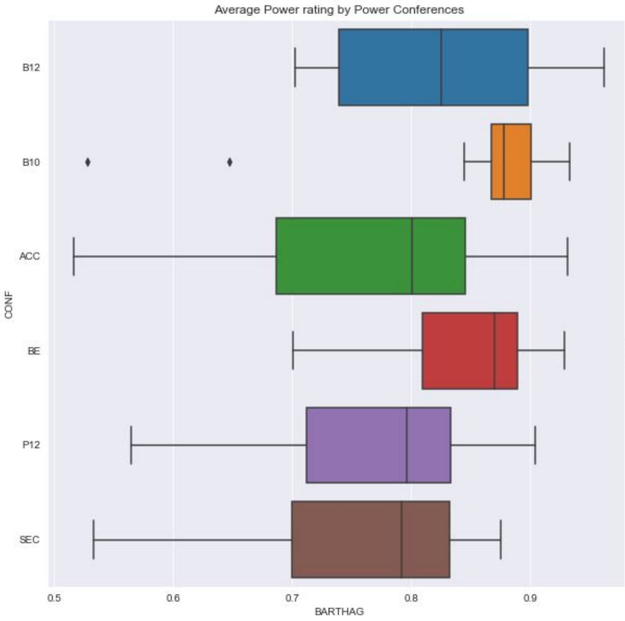
Outcome

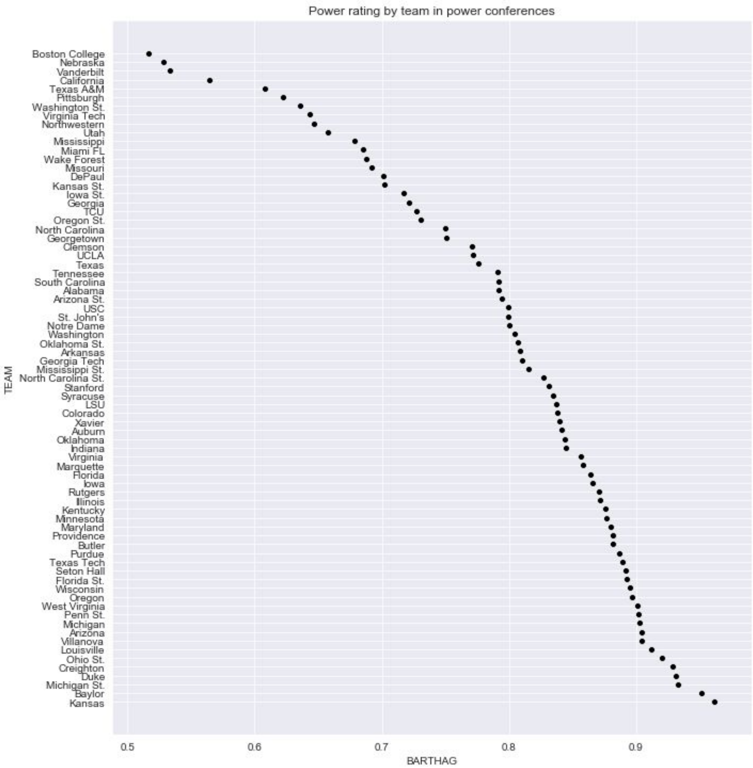
This graph is the predicted wins it takes to make the NCAA tournament. The cutoff is zero but is not limited to zero. By this graph the estimated wins it takes to make the tournament is about 23 wins.



Outcome

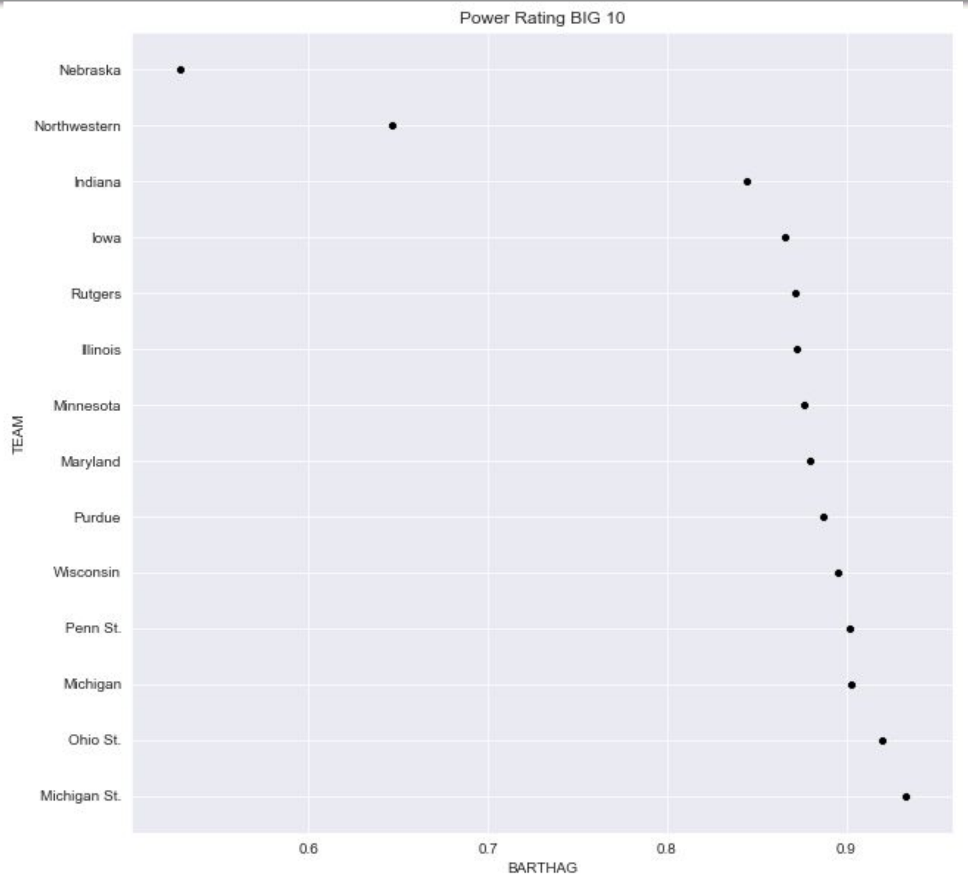
These graphs show the Power Ratings by power conference (left) and by team in those conferences (right).





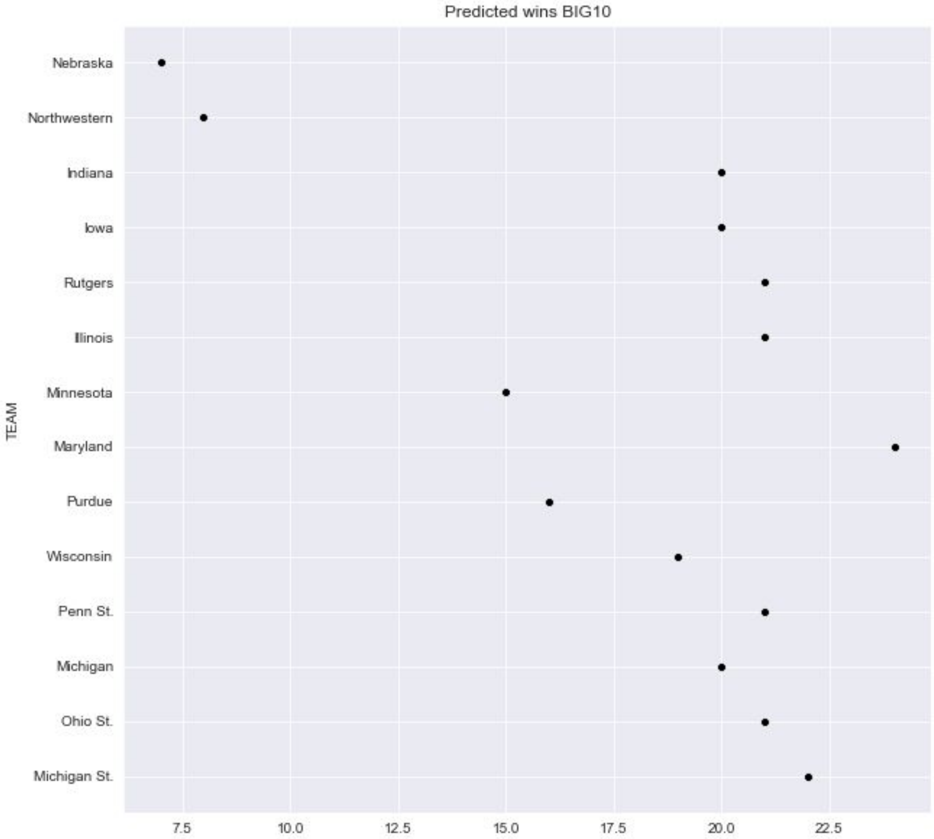
Outcome

Power rating just in the BIG 10.



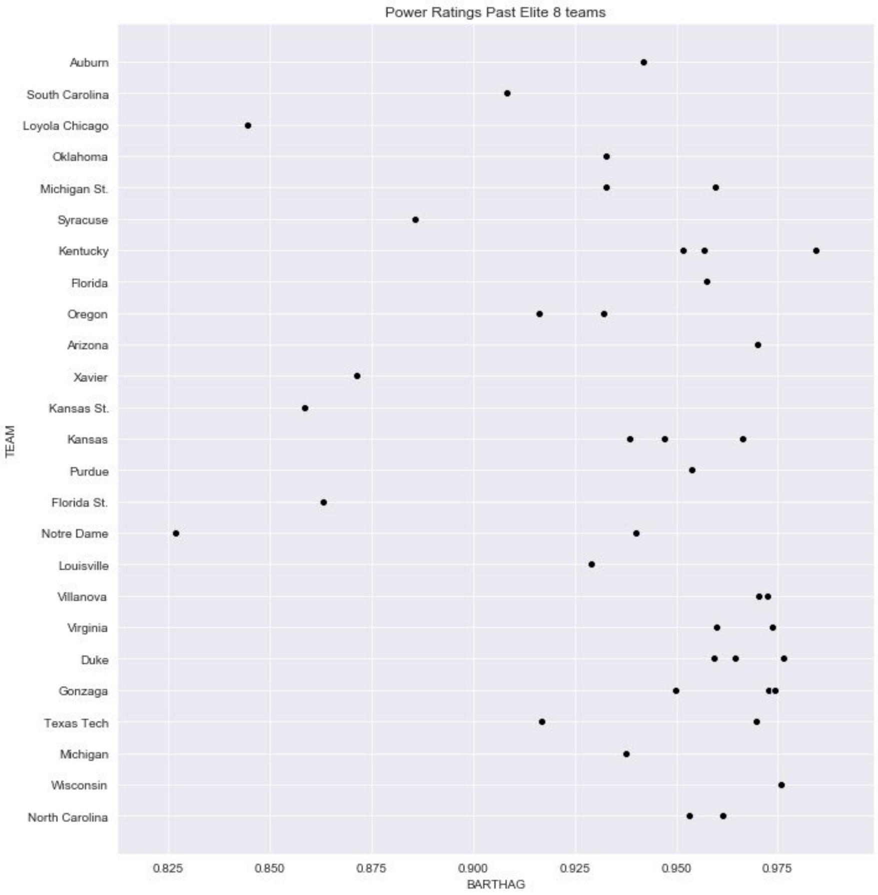
Outcome

Predicted wins for BIG 10 teams.



Outcome

Power rating for teams who have made the Elite 8 or further in the past 6 years.



Conclusion

- There are many factors in sports that are difficult to calculate that will affect wins and losses such as: luck, coaching ability, a team’s ability to handle pressure, individual player injuries, and players or teams simply having “off nights”.