Final Report: Google Cloud & NCAA® ML Competition 2019-Women's

Question

Using regular season and NCAA tournament data and external data, can we determine specific features that most influence a team's performance throughout the season? Using these top features, can we predict the March Madness bracket with some baseline accuracy?

Hypothesis

We hypothesize that certain features, such as key differences between home and away teams, region that the team is from, win percentage, distance traveled to play a game, and the amount of overtime and days of rest, have the most influence on a team's performance throughout the season.

Literature Review

Betting on the men's and women's NCAA basketball tournament has become increasingly popular over the past twenty years. For casual betters, it has been shown that bettors often overestimate how successful the favored teams going into the tournament will be (Metrick). While bettors in a larger "betting pool" overestimate the favorites slightly less, this bias is still present.

Research into building models to predict the results of the NCAA basketball tournament typically isolates two factors: which variables and statistical models should be utilized. While there is no common consensus on the best features or models to employ, statisticians have proposed many variations.

One of the most commonly used models, simple regression has been used in predicting NCAA results with success. A statistical analysis of simple regression models has shown there exists a strong correlation between a teams seed and their margin of victory across the tournament (Smith et al.). In NCAA basketball tournament predicting competitions, such as the Kaggle competition, regression models have been shown to be in the top 5% for model accuracy.

More complex models, however, are often favored by researchers. In a paper published by Kvam and Sokol, it was shown that their model, which combined logistic regression and a Markov Chain model, was successful in predicting NCAA tournament outcomes. In the past 6 years, this model outperformed the predictions put out by ESPN, USA Today and AP even when fitted with basic input data.

In research by Neudorfer et al., their model combined two methods: the least squared pairwise comparison model and isotonic regression. Using team rankings as the feature of their model, they looked for monotonic, non-linear relationships between a team's rank and it's probability of winning a game. Using simulations, their model produced better results than the commonly used linear and logistic regression model, even when the same variables were used.

In addition to research on which statistical models to choose, studies have investigated which variables potentially indicate success. Prior to the NCAA tournament, teams in each region are seeded 1-16. While logically it seems that seed and success in the tournament should be negatively correlated, it was found that seeds 10,11, and 12 have advanced to the Round of 16 a significantly greater proportion of times than seeds 8 and 9 have.

Introduction of Data (with Data Limitations)

The majority of the data used in our analysis was provided directly by the Kaggle competition page. The data is divided up into eleven datasets, each containing unique information about individual game statistics and locations and aggregated team metrics for NCAA women's basketball. Since the majority of the game data dates back to 2010, we limited all our game data to games played between 2010 and 2018 for our analysis.

The first two datasets, labeled "Cities" and "Game Cities", contain geographical data information. Cities contains the City ID for each city in which games were played while Game Cities contains the game information (i.e., Season, Game Day Number, Winning Team, Losing Team) and the city in which games were played. To calculate distances, latitude and longitude data, obtained from Simplemaps.com, were added to Game Cities. The Cities table was unused in our analysis.

The next set of tables, labeled "Tourney Compact Results", " Tourney Detailed Results", "Tourney Seed" and "Tourney Slots", contain information on the past NCAA tournaments. The table labeled Tourney Compact Results contains variables which include Winning Team, Winning Score, Losing Team, Losing Score and Winning Location. Tourney Detailed Results contains the same information included in Tourney Compact Results but is supplemented with additional game metrics which include the number of two and three point baskets made and the number of steals, fouls committed and assists for each team. Tourney Seed contains each team's seed in the NCAA tournament and the Tourney Slots contains the seed matchups for each region in the tournament.

The datasets labeled "Regular Season Compact Results" and "Regular Season Detailed Results" are similar to "Tourney Compact Results" and "Tourney Detailed Results", respectively, but contain information on the NCAA regular season games. As much of our analysis focuses on how regular season metrics can predict NCAA tournament success, these two datasets are two of the most utilized.

The "Season" dataset contains the following variables: Season, DayZero (the date of the start of the NCAA tournament), RegionW, RegionX, RegionY and RegionZ. The region variables denote the locations of where each region (i.e., West, East, Midwest, Mideast) played their tournament games before the Final 4 series, during which the four remaining teams play at a neutral location.

The remaining two datasets are labeled “Teams” and “Team Spellings”. The Teams dataset contains the team name (i.e., college name) and team ID, and the Team Spellings table contains the various abbreviated spellings of team names (e.g., A&M-Corpus Chris , A&M-Corpus Christi).

In calculating the distances a team traveled, we used geographical data from Simplemaps.com, an external data source. It is unknown if latitude and longitude data from other sources would agree with that from Simplemaps. This potential difference in geographical data could lead to differences in calculated differences, which could affect our final predictions.

In addition, another data limitation is the inconsistent presence of data from before 2010. When looking at game data, which is present in the Regular Season Compact and Detailed Results datasets and in the Tourney Compact and Detailed Results datasets, game data from before 2010 is only included in the compact datasets. As a large portion of our analysis deals with metrics only present in the detailed datasets, we decided to drop any data from before 2010 in the compact datasets and only analyze post 2010 data. While the quantity of data from post 2010 NCAA seasons is large, there could be trends in the pre-2010 data that would be valuable to the predictive power of our final model.

In the tourney_detailed dataset, the game dates in this dataset are expressed in relative terms, as the number of days since the start of the regular season, and aligned for each season so that day number #133 is the Monday right before the tournament, when team selections are made. During any given season, day number zero is defined to be exactly 19 weeks earlier than Selection Monday, so Day #0 is a Monday in late October or early November such that Day #132 is Selection Sunday (for the men's tournament) and Day #133 is Selection Monday (for the women's tournament).

This doesn't necessarily mean that the regular season will always start exactly on Day #0; in fact, during the past decade, regular season games typically start being played on a Friday that is either Day #4 or Day #11, but further back there was more variety.

To analyze the distance that each team travelled for each game, we had to obtain additional data from the Integrated Postsecondary Education Data System, or IPEDS. Specifically, we obtained data about the institution name associated with the winning and losing teams as well as each institution’s latitude and longitude coordinates, or lat-long pairs.

In terms of data limitations, the given data prior to 2010 is inconsistent, specifically the data prior to 2010 is only included in the compact datasets, but not in the detailed datasets. In addition, the given datasets do not include location data (specifically, GeoJSON data), so we need to find external datasets with this location data included in them. To resolve this issue, we had to use the us_cities dataset to get the lat-long pairs (coordinates) of the cities and states that were locations of the NCAA tournament games. However, the us_cities dataset did not include

Notre Dame, IN, and University Park, PA, so we found the lat-long pairs of those cities from additional resources.

Background

The National Collegiate Athletic Association (NCAA) is an organization which regulates collegiate athletics. NCAA sports, among others, include: basketball, soccer, football and baseball. In men's and women's NCAA basketball, teams are divided up into Division 1, Division 2 and Division 3, with Division 1 being the highest athletic level.

Every March, at the end of the NCAA basketball regular season, the NCAA basketball tournaments. The tournaments are consists of 64 teams, 16 from each of the Western, Eastern, Midwestern and Mideastern regions. Within each region, teams are seeded 1-16. In the first round, seeds 1 and 16, 2 and 15, and so on play each other. As one would expect, seeding plays a major factor in the success of a team; only one 16thseed has ever beaten a 1stseed in the NCAA men's tournament. Similarly in the 138 matchups between 2ndand 15th, only 8 15thseeds have won.

While the 1stvs 16thseed matchups might be easy to pick, no one has ever picked a perfect NCAA basketball tournament bracket. While estimating the probability of a perfect bracket isn't an exact science, Duke professor Jonathan Mattingly predicts the probability of picking a perfect bracket, accounting for predictable matchups like the ones above, to be one in 2.4 trillion.

Nevertheless, each March nearly $3 billion is bet on the NCAA basketball tournaments. Each year challenges, such as the $1 million a year for life challenge put forth by Warren Buffet, are put forth which reward perfect brackets with millions of dollars. For the last 6 years, Kaggle has hosted a "March Madness" competition where participants forecast the outcomes of the women's NCAA women's Basketball Tournament. Participants are provided with datasets containing information from past women's NCAA basketball season and are challenged to build a predictive model. This model must, for each possible tournament matchup, produce the probability of each team winning the game. Submissions are scored on the log loss:

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})\right]$$

where:

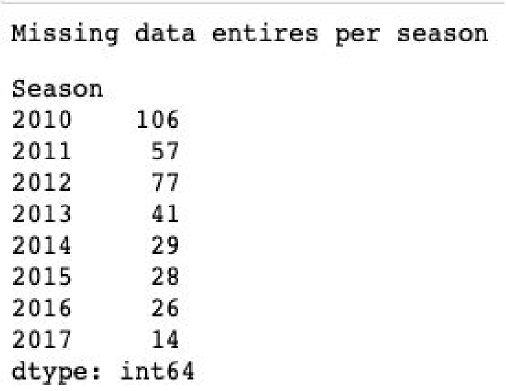
- n is the number of games played
- $\hat{y}_i$is the predicted probability of team 1 beating team 2
- yi is 1 if the team wins, 0 if team 2 wins
- log() is the natural (base e) logarithm

The team with the best predictive model wins $10,000.

Basic Analysis

Home-Court Advantage

As a part of Scenario 1 we will be looking at the performance of teams both at home and away from home. For our analysis we will just be analysing data from 2010 to 2017. Firstly, we split our data into two dataframes, one for the regular season and one for the knockout rounds. We will then merge the information about the cities the games were played in from 'merged_cities' and the compact information about the games, for both regular season and the knockout rounds. On viewing the shape of the two dataframes, we can see that we have complete information for the knockout rounds, but not for the regular season. For the regular season, we have more information in the compact results than the cities the games were played in. We are missing around 400 rows for the regular season, this may seem like a lot, but in reality that is just 0.9% of our entire dataset. Further, we are missing a maximum of 100 games in a season out of 5000, which is just about 2% of the games per season.

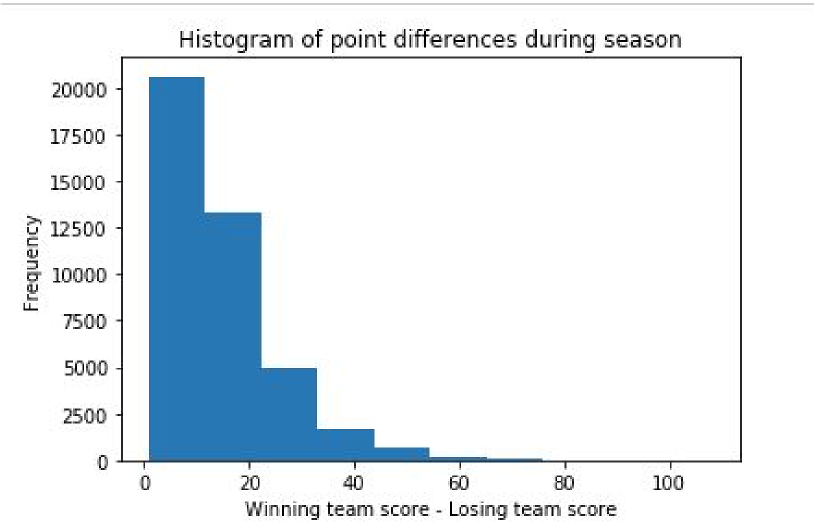


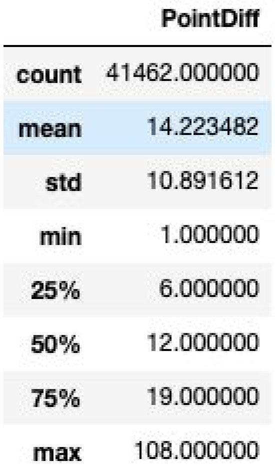
```
Missing data entires per season

Season
2010    106
2011     57
2012     77
2013     41
2014     29
2015     28
2016     26
2017     14
dtype: int64
```

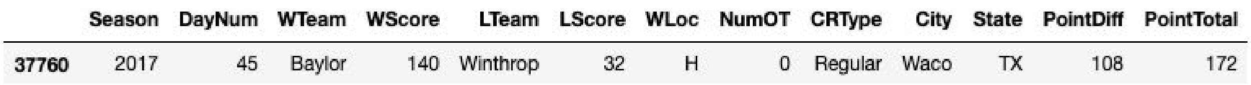
We then got the point differences between the winning and the losing team. We expect the distribution to be right skewed, since most games would be close match ups. Further, teams will play all their games seriously as their performances affect their qualification to the knockout rounds and their potential seed in the NCAA.

For our regular season we can see that the distribution is as expected. Further, the minimum point difference is 1, which is consistent with our domain knowledge, since the game would go into overtime if scores were tied. Our max score difference is however 108, which is very suspicious, but on viewing the row in the dataframe we can see that this was not a data entry error.





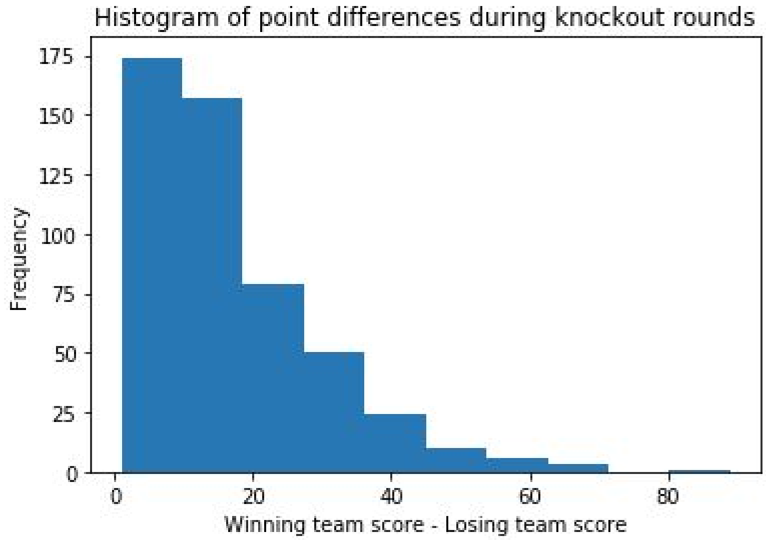
	PointDiff
count	41462.000000
mean	14.223482
std	10.891612
min	1.000000
25%	6.000000
50%	12.000000
75%	19.000000
max	108.000000

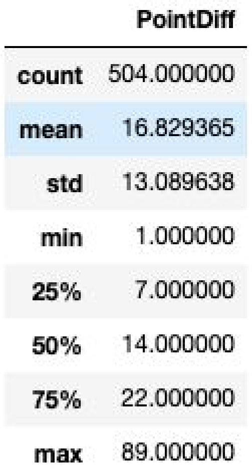


	Season	DayNum	WTeam	WScore	LTeam	LScore	WLoc	NumOT	CRType	City	State	PointDiff	PointTotal
37760	2017	45	Baylor	140	Winthrop	32	H	0	Regular	Waco	TX	108	172

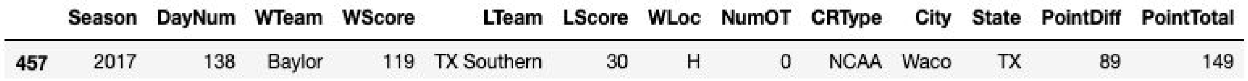
The above histogram is unimodal, with a mode of around 5. The distribution has a long right tail, making it right skewed.

For the knockout rounds, we would expect the point differences to be lower. This is because the best teams in the country are competing against each other. Further, the teams that are not as good were knocked out in the regular season. We can still however see a maximum point difference of 89, again this could potentially be an error. But on viewing the row of the table, we can see that the game was on day 138 which is the first day of the knockout rounds. Further, we know that the selection of teams for the round of 64 are picked by region (16 teams per region). This highlights that some regions may not be as competitive as others. This extra knowledge helps us understand the large point differences.





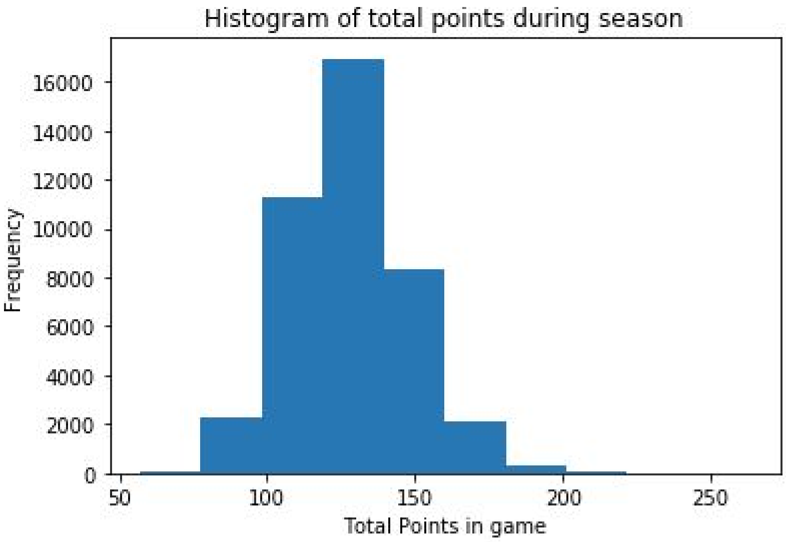
	PointDiff
count	504.000000
mean	16.829365
std	13.089638
min	1.000000
25%	7.000000
50%	14.000000
75%	22.000000
max	89.000000

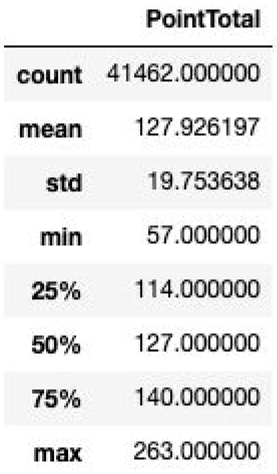


	Season	DayNum	WTeam	WScore	LTeam	LScore	WLoc	NumOT	CRType	City	State	PointDiff	PointTotal
457	2017	138	Baylor	119	TX Southern	30	H	0	NCAA	Waco	TX	89	149

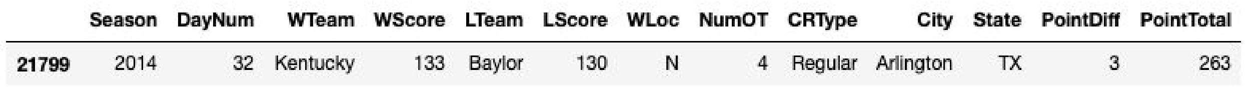
The above histogram is unimodal, with a mode of around 5. The distribution has a long right tail, making it right skewed.

We will next plot histograms for the total number of points scored in each game. We would expect a normal distribution for this, as we are less likely to have extremely high scoring and low scoring games. Again we will generate two different plots for the regular season and NCAA. On viewing the average statistics, we can see that the mean is very close for regular season and NCAA. The minimum for regular season is a lot lower than NCAA, this could be because the quality of the game is a lot higher in the NCAA, with only the best 64 teams in the country qualifying for NCAA. The maximum points were scored in the regular season, this could be because the teams were very good offensively, but not defensively. These teams with just a good attack and weak defense would not do well in knockout rounds as you need a balance of attack and defense in your team to succeed in the later rounds of the competition.

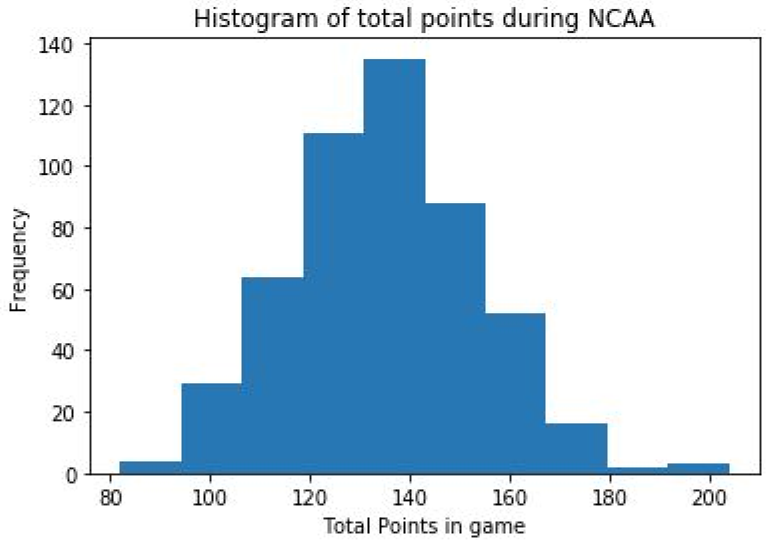


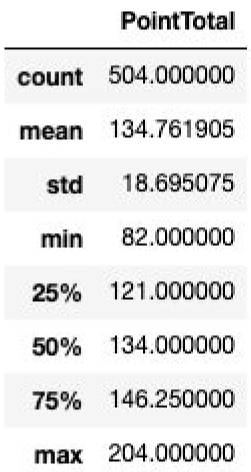


	PointTotal
count	41462.000000
mean	127.926197
std	19.753638
min	57.000000
25%	114.000000
50%	127.000000
75%	140.000000
max	263.000000



	Season	DayNum	WTeam	WScore	LTeam	LScore	WLoc	NumOT	CRType	City	State	PointDiff	PointTotal
21799	2014	32	Kentucky	133	Baylor	130	N	4	Regular	Arlington	TX	3	263



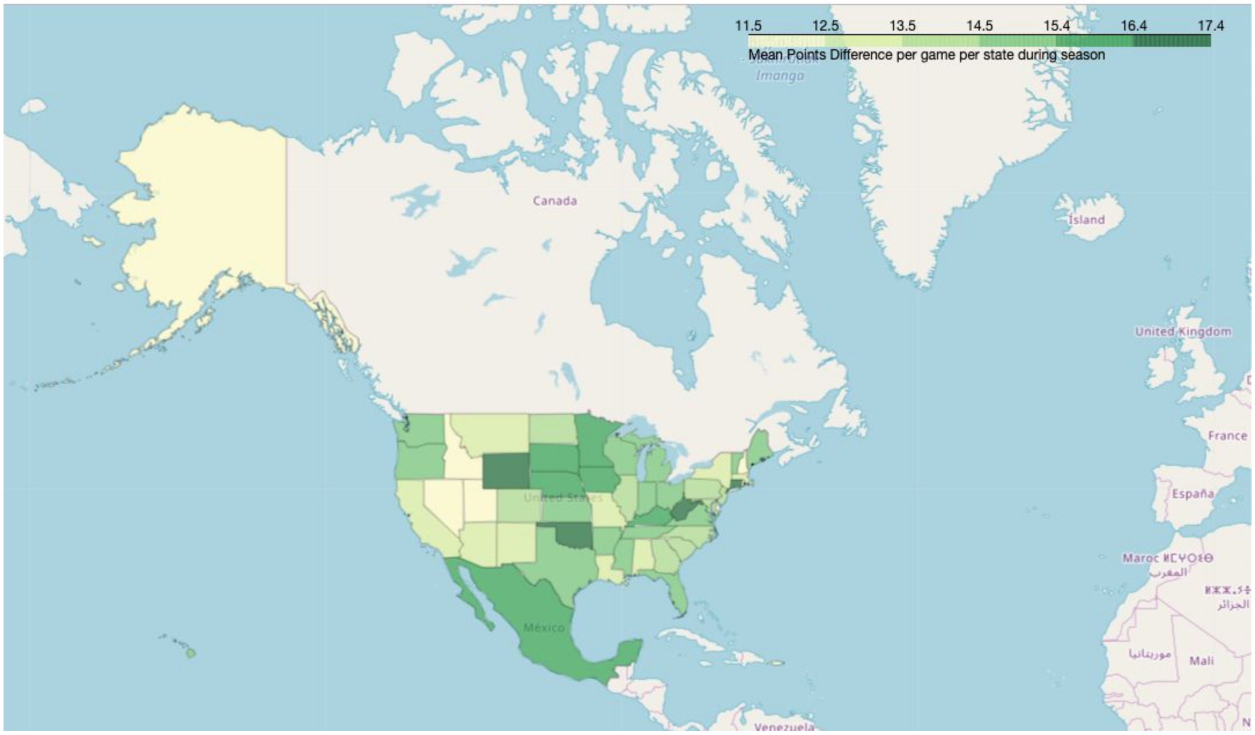


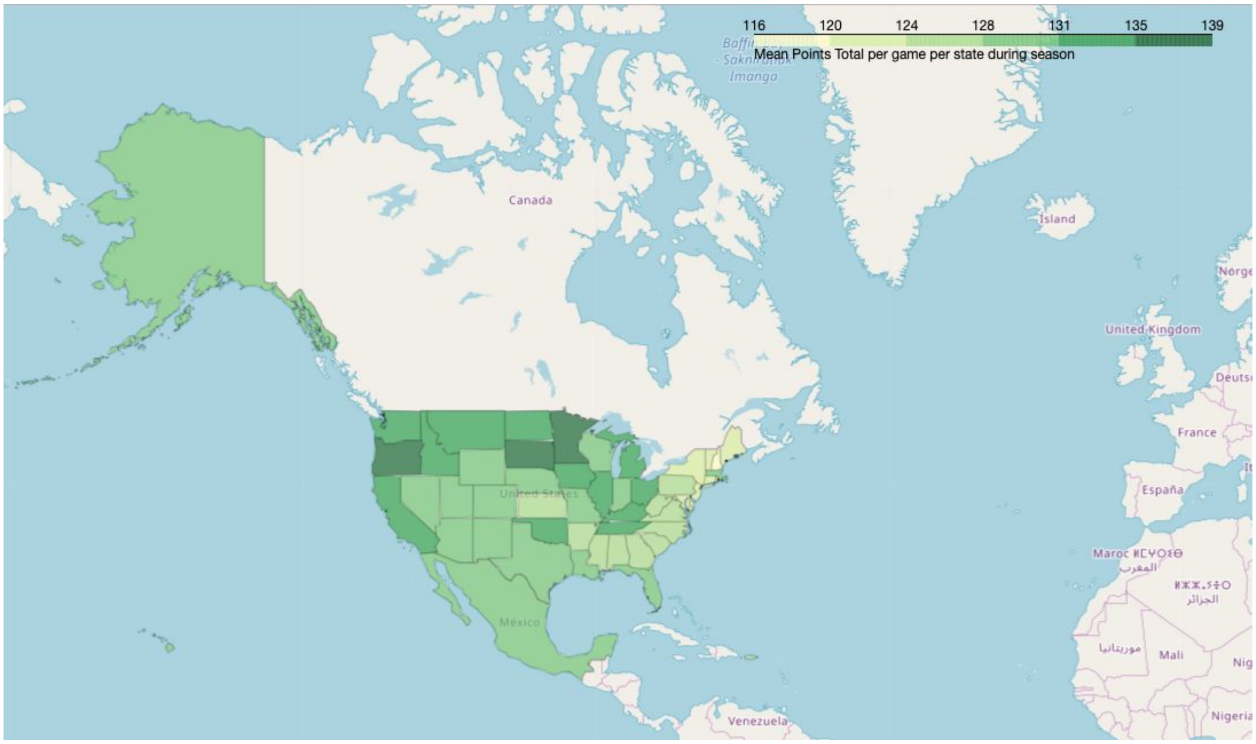
	PointTotal
count	504.000000
mean	134.761905
std	18.695075
min	82.000000
25%	121.000000
50%	134.000000
75%	146.250000
max	204.000000

The above histograms are unimodal and symmetric. The first histogram has a mode of around 130, while the second histogram has a mode of 135.

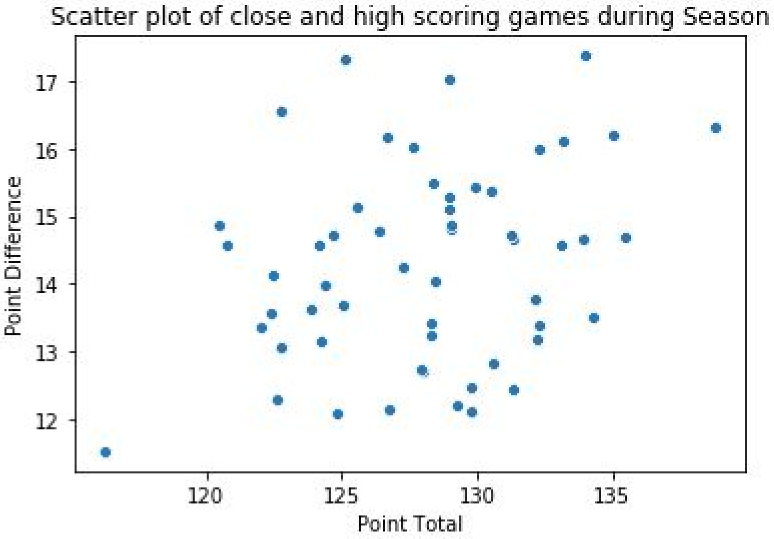
In the next part of our analysis, we will look at which states have the reputation of producing high scoring games, and which states produce the closest match ups for both the regular season and for the NCAA. We will use choropleths to visualize the data. We must note that we could not get the GeoJSON files for the states of Ontario and Bahamas.

For games during the season Ontario has the highest total point average. But on viewing the data, we can see that there was just one game played in Ontario. So it is for the best that we don't include the state 'ON' in our choropleth and we will drop the row from our dataframe. In general we can see that Minnesota and Oregan has higher scoring games, while cities in the north east (New Hampshire, Maine etc) of the US have lower scoring games. On plotting the point differences, we can see that states like New Hampshire and Rhode Island have closer games, while games in Oklahoma and West Virginia are more one sided. On first glance, it seems that lower scoring games tend to be close results. We plotted our findings on a scatterplot and can see that the correlation between point difference and point total is around 0.3, which indicates that there is a weak positive correlation.

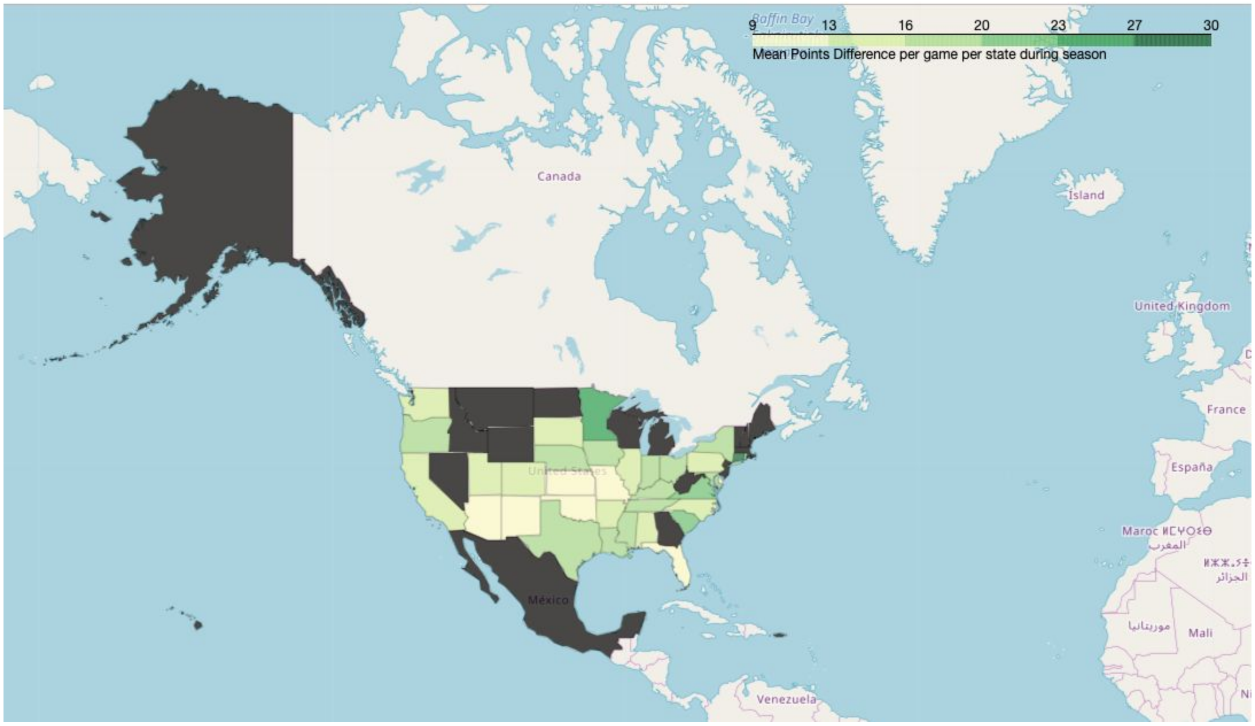


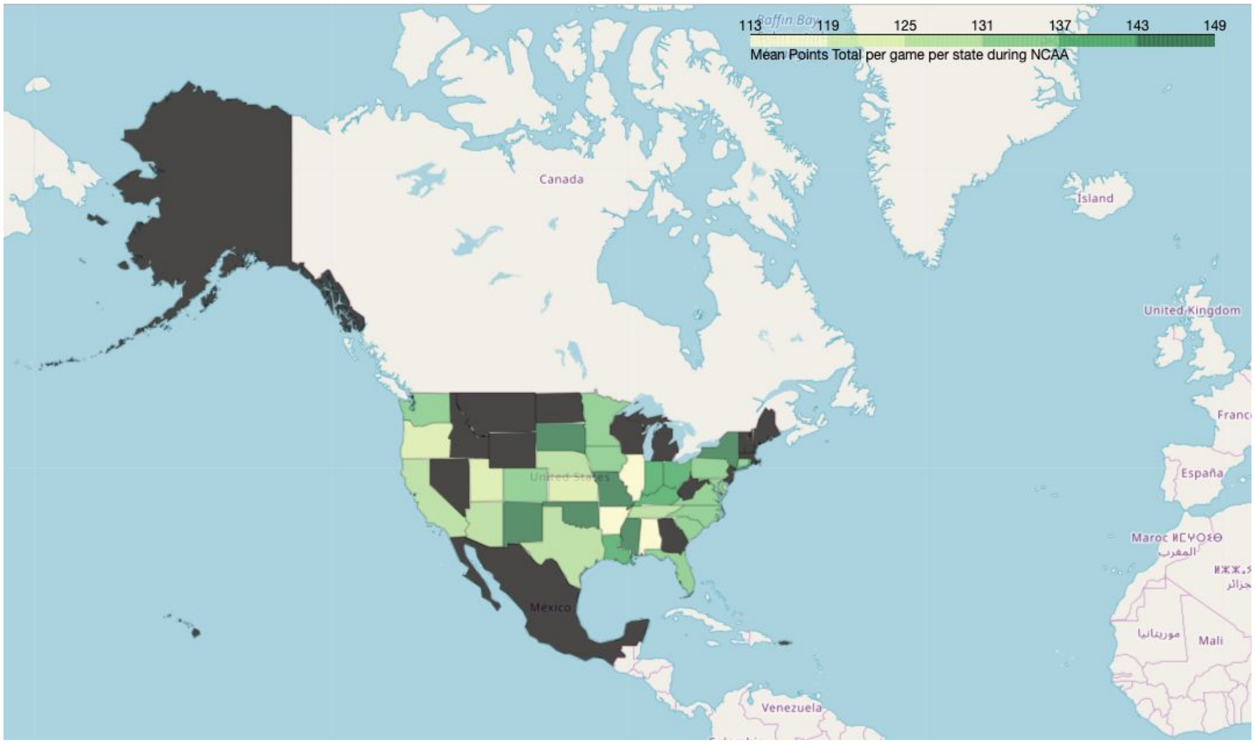


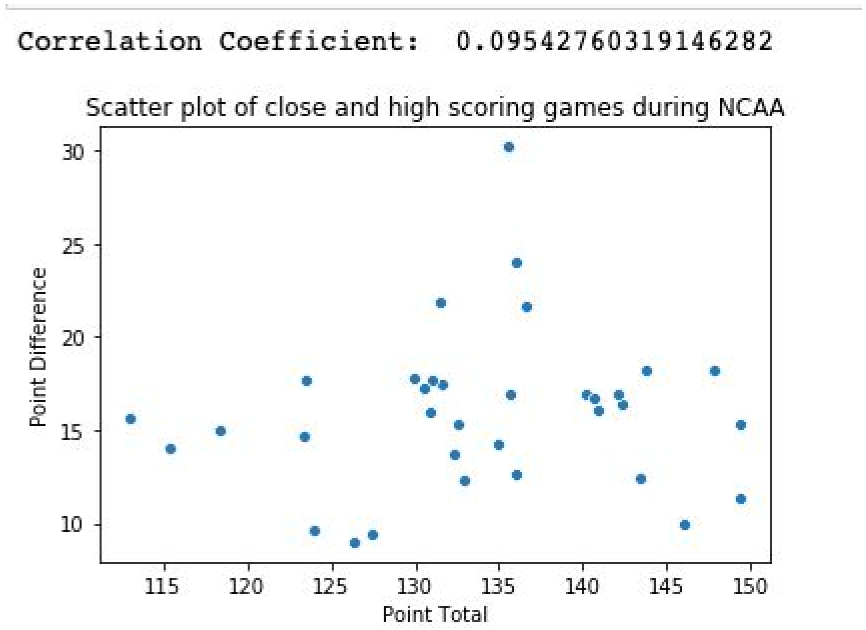
```
Correlation Coefficient:  0.28481299791104614
```



Next we will look at how the same information during the NCAA. On viewing the size of each group we can see that there are a lot of states with very few matches. We can see that 11 states have only had 3 matches each. Most states do have a high point average for the NCAA, this is probably because 2/3rds of the games are played on neutral grounds. Further, we have a lot of states in the US that have not had any NCAA games. On first glance of the choropleths, we can see that the states that have higher point totals (New Mexico, Missouri) have lower point differences. This is the ideal game for viewers to watch, high scoring and close. On studying the correlation between the two variables, we could however see that those trends were specific to those two states and not the others.

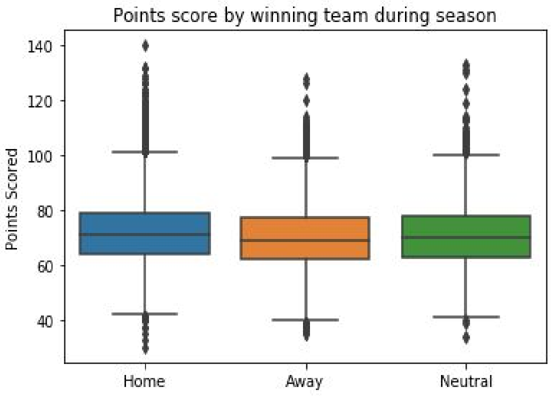


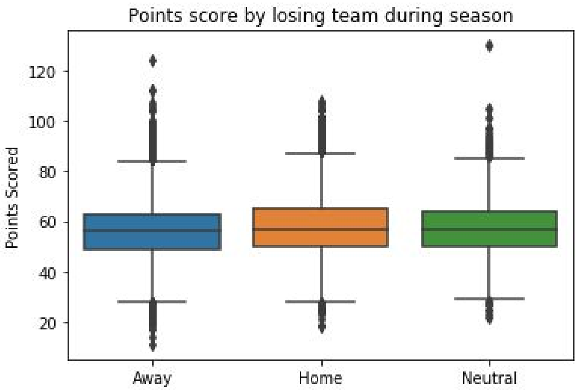




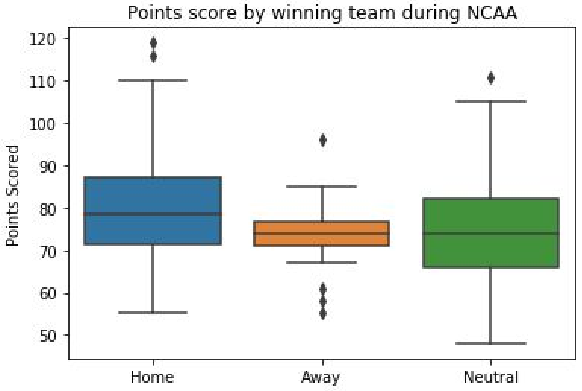
```
Correlation Coefficient:  0.09542760319146282
```

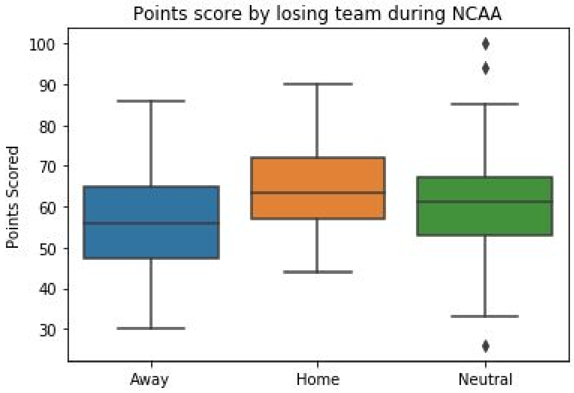
Next, we will look into what a team has to do to win a game away from home. We will do that by plotting boxplots of the winning scores, losing scores based on if the team was home or away. This may give us an indication of whether the away team should concentrate on defense or attack on their travels. The boxplots for the points scored by the teams during the season is almost identical. There is no indication of whether attack or defense is more important.





But when we plot the same boxplots for the NCAA, we can see that home losing sides have significantly more points scored than away losing sides. This means that when a home team loses, they still score a lot of baskets, and if an away side wants to win away from home, they must concentrate on their attacking play. Though the defensive play is important, their offense is the differential between the away team winning and losing. Our sample for teams winning away is pretty small for the NCAA for two reasons, one being that most games are played on neutral territory and the second being that the home support would usually carry their team over the finish line.

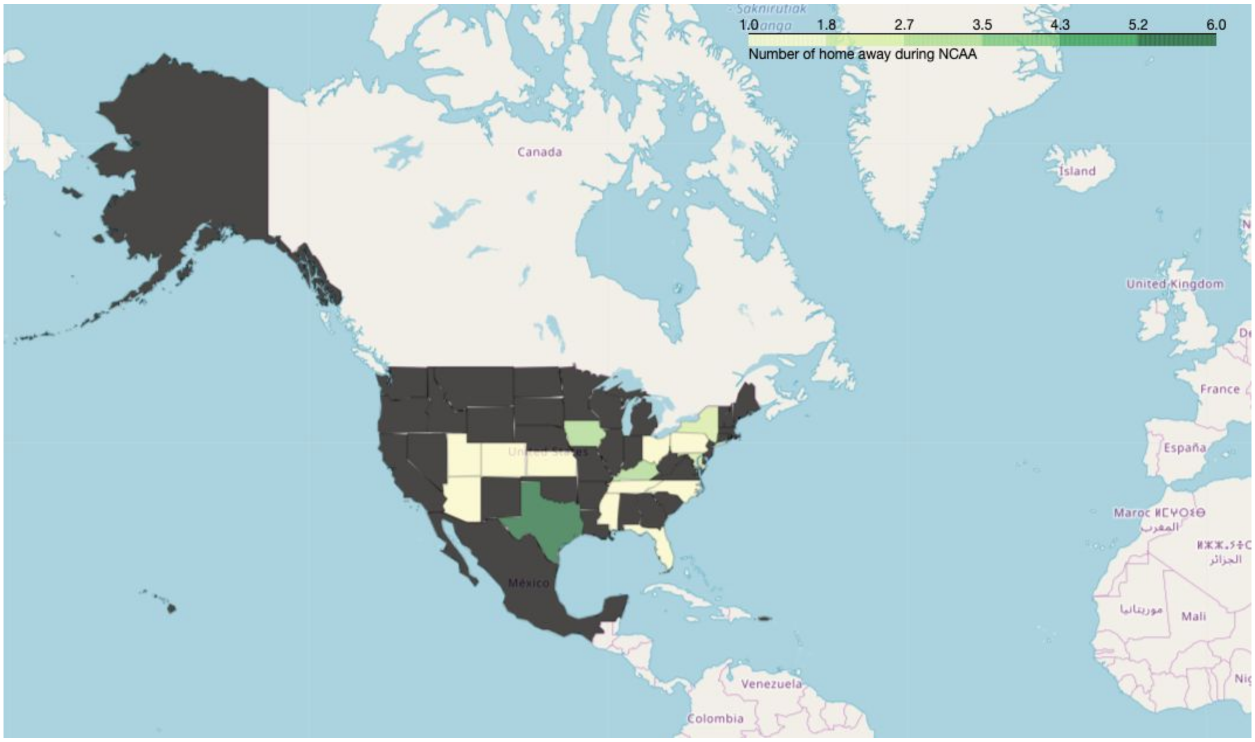


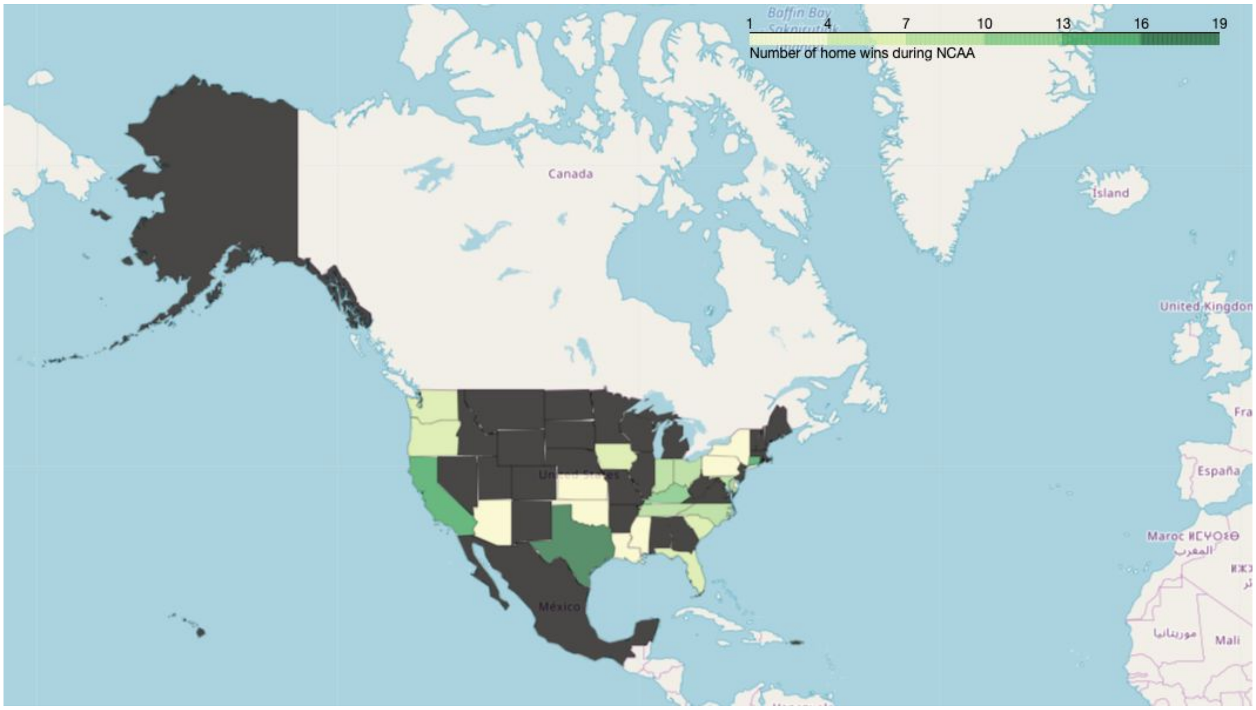


For the next step of our analysis, we will look at which state tends to be a home fortress for their teams and which state has stadiums with grounds where away teams usually come out victorious. This may give us an indication of how passionate the people from that state are with regards to women's basketball. Having a supportive home crowd will usually help push you over the finish line in times of difficulty.

It wouldn't really make sense for us to analyse data during the season, since the divisions are split by region. So you will mostly play other schools from the state. if we did plot this data, we would expect heavily populated states to have more home and away wins, solely because you will see more matches where both teams are from the same state.

We can however gather more information, if we use a choropleth to analyse data for teams that travel during the NCAA. For example, we can see that away teams often win in Texas. If we plot the same choropleth for home wins, this time round we have California joining Texas. It is odd that Texas makes both the lists for state with a lot of home wins as well as away wins. This indicates that Texas has more teams qualifying for the NCAA and have a history of reaching the later rounds in the tournament, leading to high number of wins and losses.

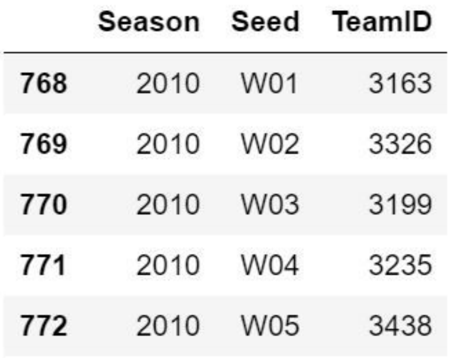

1.0
1.8
2.7
3.5
4.3
5.2
6.0
Number of home away during NCAA
Canada
Ísland
United Kingdom
France
España
Maroc
المغرب
الجزائر
موريتانيا
Mali
United States
México
Venezuela
Colombia


1
4
7
10
13
16
19
Number of home wins during NCAA
Baffin Bay
Canada
Ísland
United Kingdom
España
Maroc
المغرب
موريتانيا
Mali
United States
México

How Predictive are the Seedings?

We also thought a team's seed might be a useful feature for developing a model that predicts tournament success once the brackets are set. Seeds are generated based on a team's performance during the regular season, using a variety of advanced statistics. These seedings determine the structure of the NCAA Tournament. Each region has 16 seeds. In each region, the 1 seed plays the 16 seed, the 2 seed plays the 15 seed, and so on. The intent of this structure is so that good teams will not be defeated early in the tournament by the best teams. Ultimately, the goal is to leave the best matchups for the latest in the tournament.

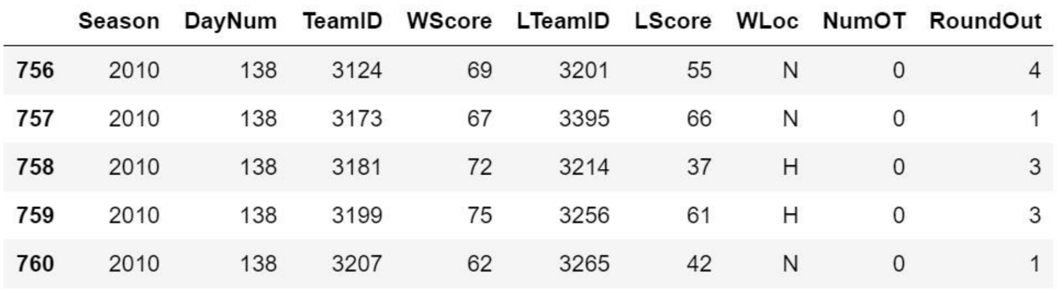
Seeds are supposed to be based on a team's expected performance in the tournament, so it is rare for a 16 seed to defeat a 1 seed or a 15 seed to defeat a 2 seed. To run association tests on this data, we first had to clean and merge data corresponding to seedings as well as tournament performance. The raw seed data looks as follows:



	Season	Seed	TeamID
768	2010	W01	3163
769	2010	W02	3326
770	2010	W03	3199
771	2010	W04	3235
772	2010	W05	3438

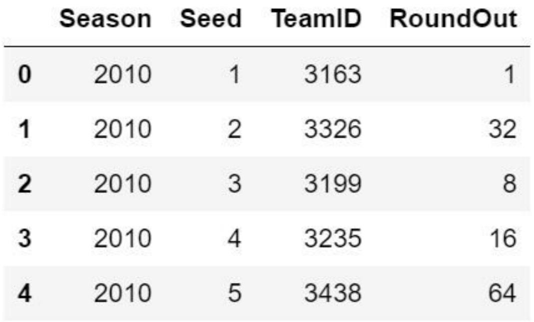
To determine tournament performance, we looked at the round at which a team was defeated (which will be referred to as "RoundOut"). To find this, we computed the number of times a team appeared as the winning team in a tournament. As such, zero would correspond to losing in the round of 64 and six would correspond to winning the tournament. Those counts were replaced by a representation of the team's "RoundOut." Zero was replaced with 64, one was replaced with 32, and so on. The raw data from which we found these tournament placings are as shown, followed by the merged data on which we performed our analysis:

Raw Data with RoundOut:



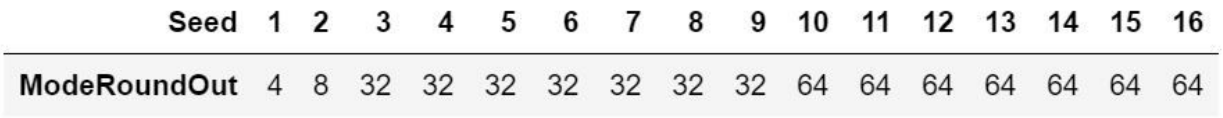
	Season	DayNum	TeamID	WScore	LTeamID	LScore	WLoc	NumOT	RoundOut
756	2010	138	3124	69	3201	55	N	0	4
757	2010	138	3173	67	3395	66	N	0	1
758	2010	138	3181	72	3214	37	H	0	3
759	2010	138	3199	75	3256	61	H	0	3
760	2010	138	3207	62	3265	42	N	0	1

Merged Data:



	Season	Seed	TeamID	RoundOut
0	2010	1	3163	1
1	2010	2	3326	32
2	2010	3	3199	8
3	2010	4	3235	16
4	2010	5	3438	64

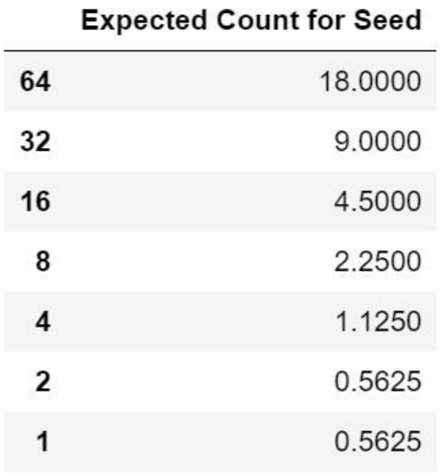
For some preliminary analysis, we thought it would be valuable to look at the most frequent RoundOut for each seed. To do this, we grouped by seed and found the mode round at which the team was eliminated. These statistics are as follows:



Seed	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ModeRoundOut	4	8	32	32	32	32	32	32	32	64	64	64	64	64	64	64

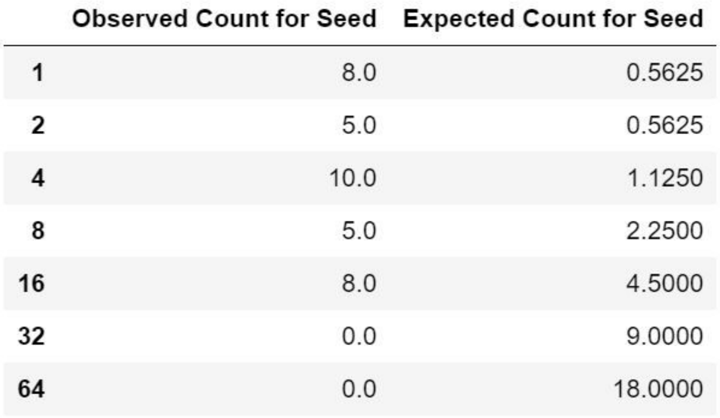
To clarify, ‘64’ means that the most frequent tournament placing for that seed is losing in the round of 64. The most frequent placing for first seeds is losing in the final four. There seems to be a clear difference in average tournament performance depending on seed, though the extent of that association is still unclear.

We then ran Chi-squared tests on expected and observed counts of each RoundOut for each seed. This process would generate a unique chi-squared value for each seed, which would express how different each observed counts were from average expected placements. Expected counts were calculated by looking at the total counts of each RoundOut and dividing by 16 to get expected counts regardless of seed. These expected counts for each round are as follows:



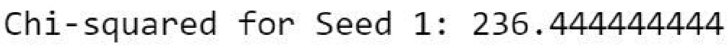
	Expected Count for Seed
64	18.0000
32	9.0000
16	4.5000
8	2.2500
4	1.1250
2	0.5625
1	0.5625

Observed counts were then generated by limiting the data to teams with a specific seed and looking at the distribution of tournament placements. The observed and expected counts for the first seed are as follows:



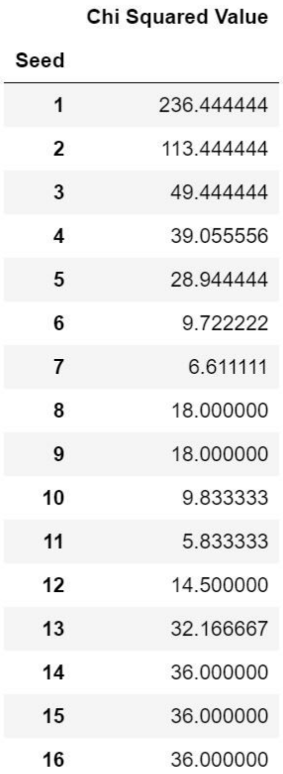
	Observed Count for Seed	Expected Count for Seed
1	8.0	0.5625
2	5.0	0.5625
4	10.0	1.1250
8	5.0	2.2500
16	8.0	4.5000
32	0.0	9.0000
64	0.0	18.0000

Running a chi-squared test on this data generates a chi-squared value for the first seed:



```
Chi-squared for Seed 1: 236.444444444
```

These values were computed for each seed, 1-16, as the results are as follows:



	Chi Squared Value
Seed	
1	236.444444
2	113.444444
3	49.444444
4	39.055556
5	28.944444
6	9.722222
7	6.611111
8	18.000000
9	18.000000
10	9.833333
11	5.833333
12	14.500000
13	32.166667
14	36.000000
15	36.000000
16	36.000000

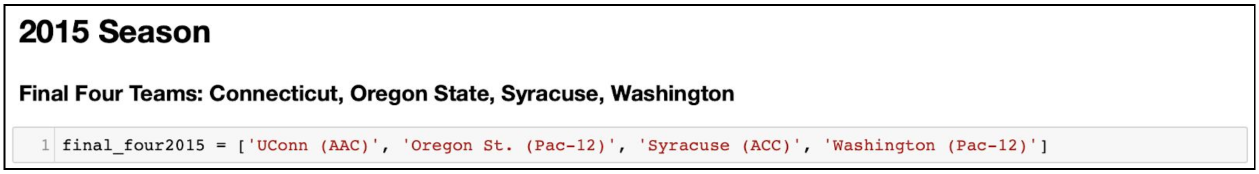
Clearly, no observed distribution of counts of each RoundOut is modeled after the uniform distribution of expected counts. It is obvious that there is an association between seed and tournament performance. It is also clear that for low and high seeds this difference is more drastic. Seed 11's performance actually looks somewhat similar to the expected counts distribution, but seeds 1-5 and 13-16 in particular look vastly different.

In our final model since we are predicting whether a team makes the NCAA tournament and, if they do, how far they will go in the tournament, we won't use seed as a variable in our final model. If we were to have the NCAA tournament brackets and were to build a model that predicted each teams success in the tournament, seed would definitely be utilized.

Advanced Metrics

For our third scenario, we took a look at the differences in the "hard-work" metrics for teams admitted to the NCAA Tournament. It has often been said that, contrary to offensive firepower, defense is an attribute that travels with teams wherever it is that they play. Therefore, we chose the following statistics to measure: Personal Fouls per Game (*PFPG*), Rebounds per Game (*RPG*), Steals per Game (*STPG*), and Blocks per Game (*BKPG*). In order to detect a statistical difference amongst attributes, we conducted various t-tests in order to decide whether or not these statistics are the difference maker when it comes to whether or not a team makes the Final Four.

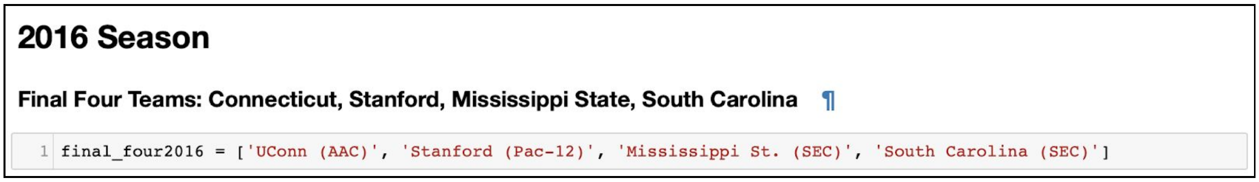
For this scenario, we limited our sample space to the last three full seasons of women's basketball in the NCAA. Contrary to professional basketball, players only spend a limited amount of time with their respective teams seeing as though students are only allotted four years of eligibility. Listed below are the teams that earned spots in the Final Four over the last three seasons (along with their respective conferences).



2015 Season

Final Four Teams: Connecticut, Oregon State, Syracuse, Washington

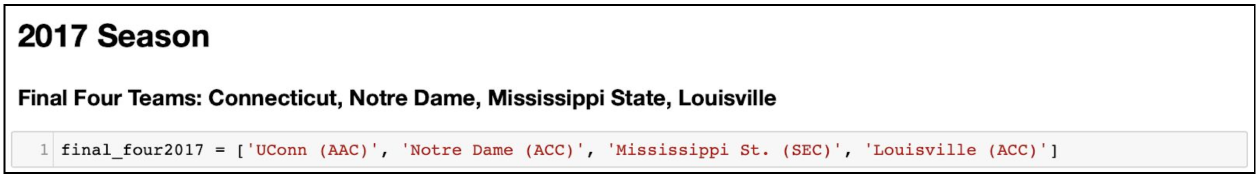
```
final_four2015 = ['UConn (AAC)', 'Oregon St. (Pac-12)', 'Syracuse (ACC)', 'Washington (Pac-12)']
```



2016 Season

Final Four Teams: Connecticut, Stanford, Mississippi State, South Carolina ¶

```
final_four2016 = ['UConn (AAC)', 'Stanford (Pac-12)', 'Mississippi St. (SEC)', 'South Carolina (SEC)']
```



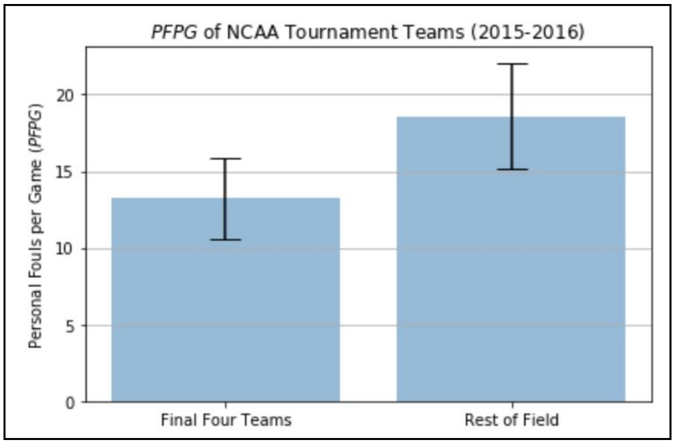
2017 Season

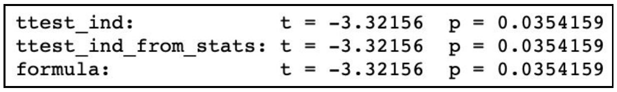
Final Four Teams: Connecticut, Notre Dame, Mississippi State, Louisville

```
final_four2017 = ['UConn (AAC)', 'Notre Dame (ACC)', 'Mississippi St. (SEC)', 'Louisville (ACC)']
```

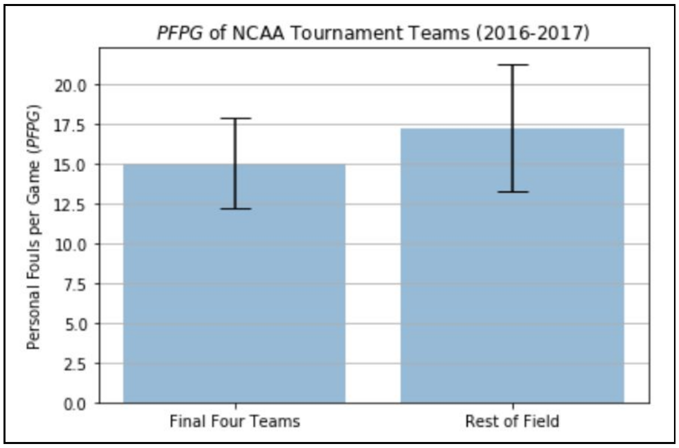
Something interesting to make note of here is that the University of Connecticut (*UConn*) has made every Final Four since 2008 (a bit of a dynasty). We will begin by looking at the first statistic of interest, Personal Fouls per Game. The reasons that we are interested in this statistic are twofold: the first is that teams that are less efficient when it comes to defensive proficiency will foul more often, and the second is that when a team reaches a certain foul threshold the other team is given the opportunity to shoot free throws and score more points.

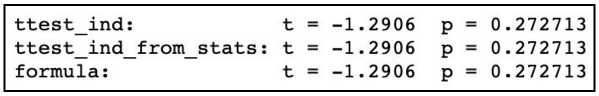
Fouls per Game



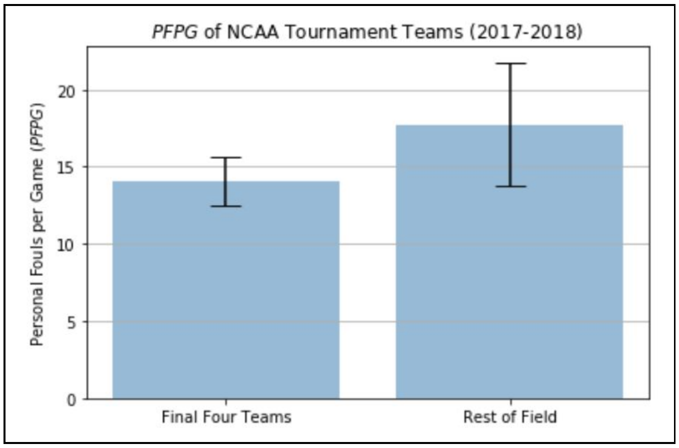


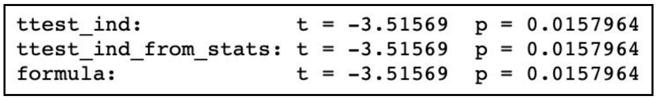
```
ttest_ind:             t = -3.32156  p = 0.0354159
ttest_ind_from_stats: t = -3.32156  p = 0.0354159
formula:               t = -3.32156  p = 0.0354159
```



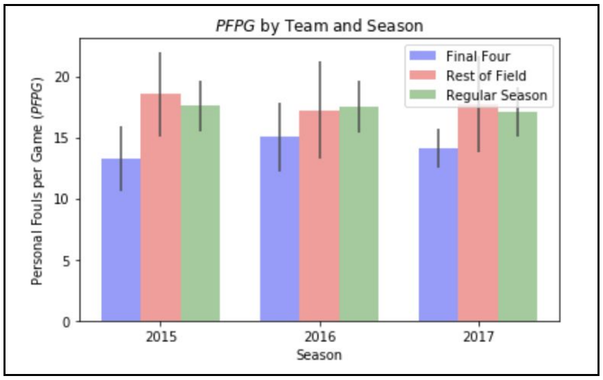


```
ttest_ind:             t = -1.2906  p = 0.272713
ttest_ind_from_stats: t = -1.2906  p = 0.272713
formula:               t = -1.2906  p = 0.272713
```





```
ttest_ind:             t = -3.51569  p = 0.0157964
ttest_ind_from_stats: t = -3.51569  p = 0.0157964
formula:               t = -3.51569  p = 0.0157964
```



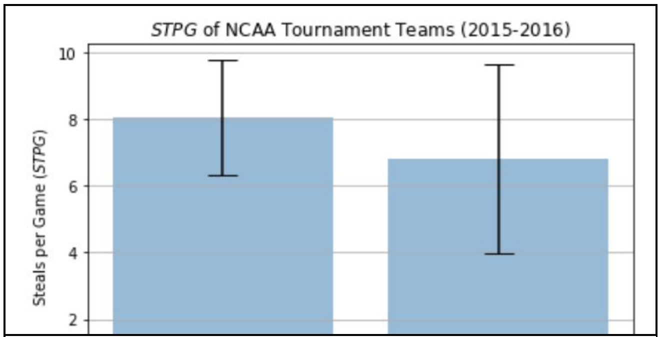
It is clear from the bar plots above that the teams that make it further in the NCAA tournament tend to pick up a lower amount of fouls than those teams that exit earlier in the tournament. On average, the teams that have made it to the Final Four the past three years have averaged *13.225, 15.025,* and *14.100* fouls per game whereas the teams that did not averaged *18.558*, *17.225*, and *17.755* in the same tournament. Using the following null hypothesis:

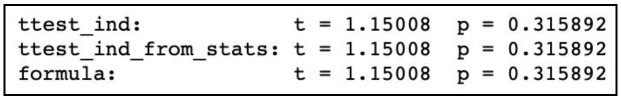
H_0: no difference in average number of fouls committed
H_A: difference exists in the number of fouls committed

and a significance level of *0.05* we see that our results have shown statistical significance in this category. We have *p-values* of *0.035*, *0.272*, and *0.016*. Therefore, we reject the null hypothesis in favor of the alternative; there exists a statistically significant difference in the average number of fouls committed amongst the final four teams and the rest of the field.

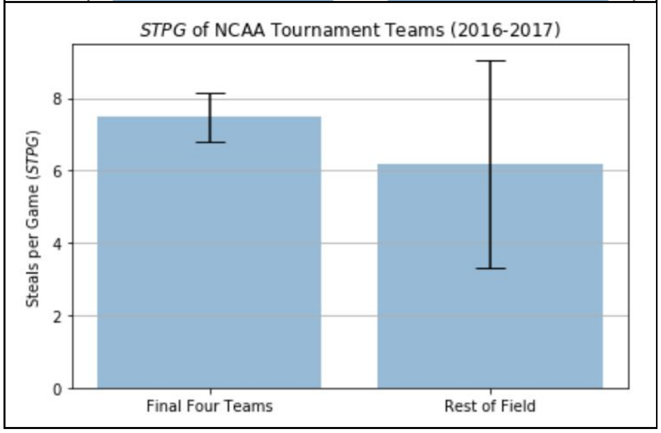
Furthermore, from the combined bar plot above we see the regular season is not a great indicator of how far a team will go in the tournament considering how many games there are in the season compared the maximum games a team can play in the tournament, six.

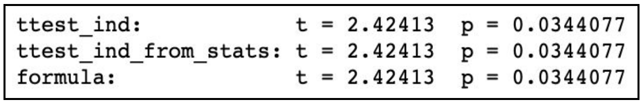
Steals per Game



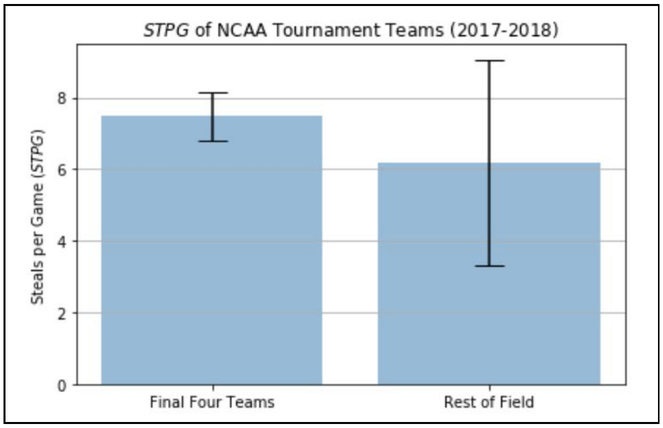


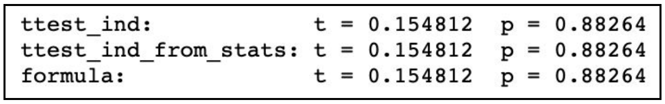
```
ttest_ind:            t = 1.15008   p = 0.315892
ttest_ind_from_stats: t = 1.15008   p = 0.315892
formula:              t = 1.15008   p = 0.315892
```



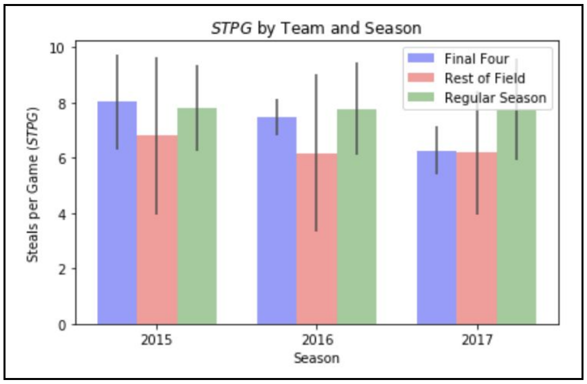


```
ttest_ind:            t = 2.42413   p = 0.0344077
ttest_ind_from_stats: t = 2.42413   p = 0.0344077
formula:              t = 2.42413   p = 0.0344077
```





```
ttest_ind:            t = 0.154812   p = 0.88264
ttest_ind_from_stats: t = 0.154812   p = 0.88264
formula:              t = 0.154812   p = 0.88264
```



The next statistic that we chose to investigate for the women's NCAA tournament was the average number of steals a team forces over the course of the tournament. Teams that have made the Final Four over the past three years have, on average, had *8.025*, *7.475*, and *6.275* steals per game whereas teams that did not earn a spot in the Final Four have had *6.803*, *6.177*, and *6.185* steals per game in the '15-'16, '16-'17, and '17-'18 seasons respectively.

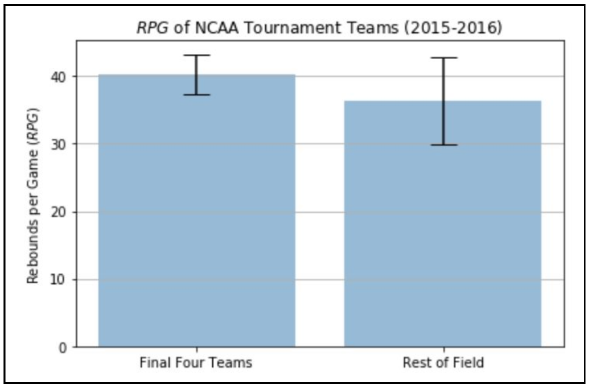
Using the following null hypothesis:

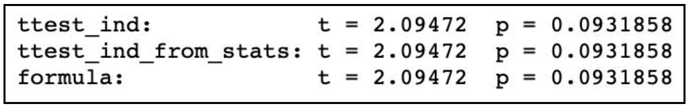
H_0: *no difference in average number of steals per game*
H_A: *there exists a difference in the number of steals per game*

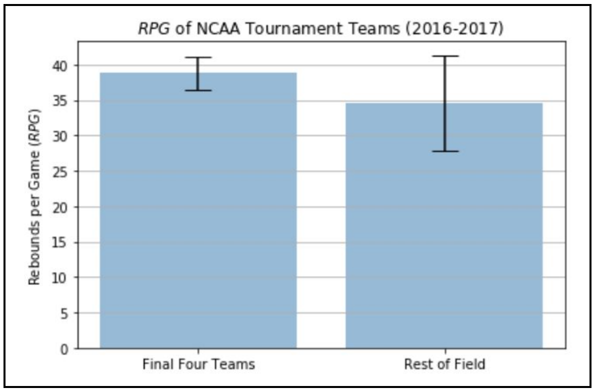
and a significance level of *0.05* we see that our results have, on average, not shown statistical significance in this category. We have *p-values* of *0.316*, *0.034*, and *0.883*. Therefore, we fail to reject the null hypothesis in favor of the alternative; a statistically significant difference in the average number of steals per game amongst the final four teams and the rest of the field does not exist. This makes sense given how small the small number of steals per game is on average. Women play four ten-minute quarters, so the average number of steals per quarter ranges anywhere from *1.544* to *2.006*.

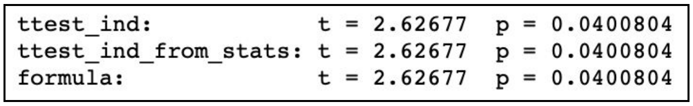
Furthermore, from the combined bar plot above we see the regular season is not a great indicator of how far a team will go based on the average number of steals per game that occur. Perhaps there are some teams that play excellent defense, however, their lack of offensive ability prevents them from winning many games. Thus, they are barred from going deep into the tournament.

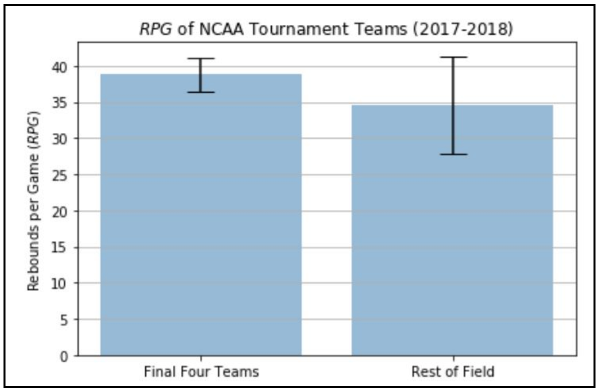
Rebounds per Game

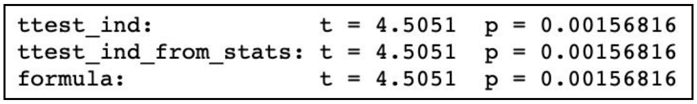


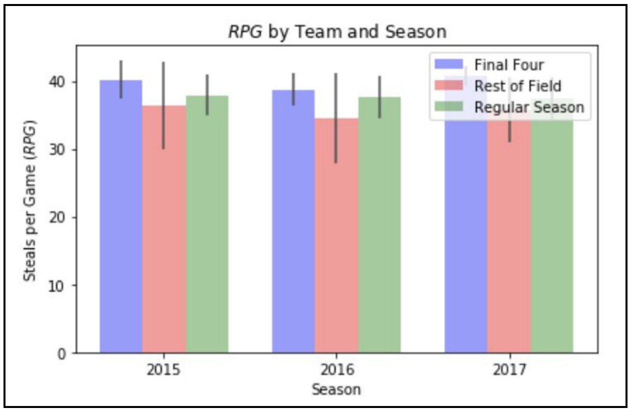

```
ttest_ind:             t = 2.09472  p = 0.0931858
ttest_ind_from_stats:  t = 2.09472  p = 0.0931858
formula:               t = 2.09472  p = 0.0931858
```




```
ttest_ind:             t = 2.62677  p = 0.0400804
ttest_ind_from_stats:  t = 2.62677  p = 0.0400804
formula:               t = 2.62677  p = 0.0400804
```




```
ttest_ind:             t = 4.5051   p = 0.00156816
ttest_ind_from_stats:  t = 4.5051   p = 0.00156816
formula:               t = 4.5051   p = 0.00156816
```



Next, we chose to look into the number of rebounds a team grabs during their games on average. Rebounds are an important statistic to measure here for a couple reasons. The ability for a team to rebound at a high rate disallows their opponent from obtaining any second chance points, and thus holds them to a lower scoring total. Secondly, if a team is able to obtain rebounds on the offensive side of the ball, they have a greater number of opportunities to put the ball in the basket and score more points.

The average number of rebounds per game for teams over the course of the season falls between our averages for the same statistics during the tournament. This fails to tell us much about how a teams performance during the season in this category affects their postseason success.

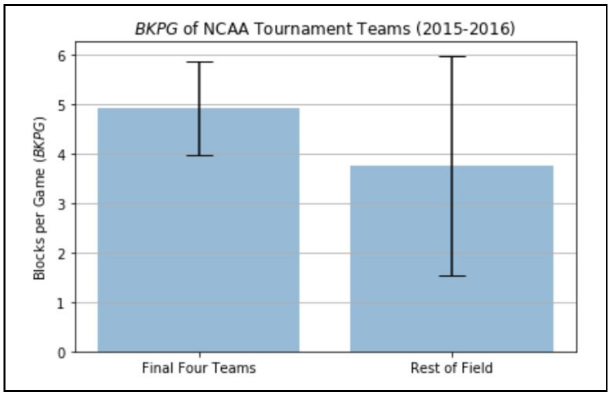
Teams that have made it to the Final Four over the last three years have averaged *40.225*, *38.760*, and *40.758* rebounds per game whereas teams that have failed to make the Final Four have averaged a lower number of average rebounds per game; *36.335*, *34.497*, and *35.768* respectively.

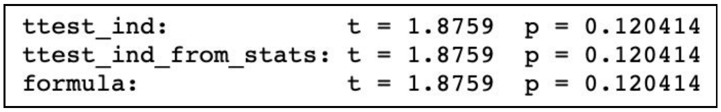
Using the following null hypothesis:

H_0*: no difference in average number of rebounds per game*
H_A*: there exists a difference in the number of rebounds per game*

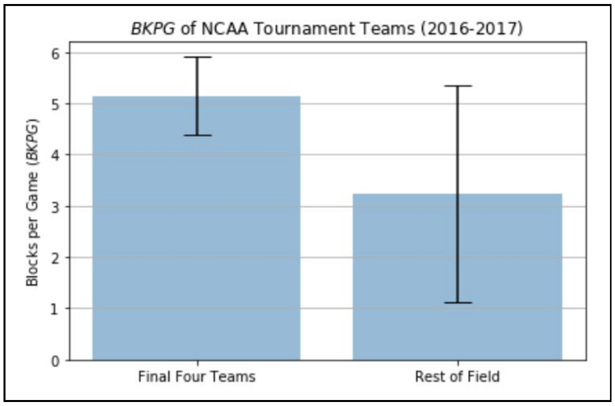
and a significance level of *0.05* we see that our results have, on average, shown statistical significance in this category. For the past three seasons we have *p-values* of *0.093*, *0.040*, and *0.002*. This tells us that teams that are able to rebound at a higher rate, on average, are able to move deeper into the tournament and have a better chance of reaching the Final Four.

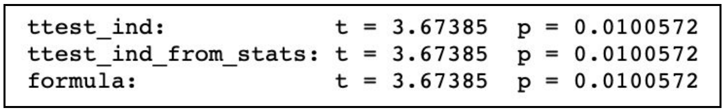
Blocks per Game



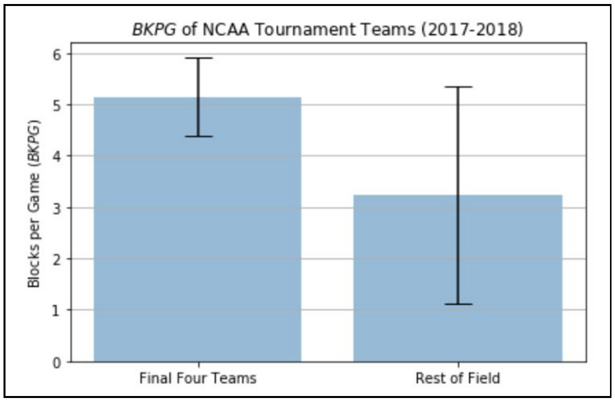


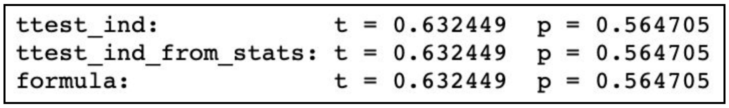
```
ttest_ind:            t = 1.8759  p = 0.120414
ttest_ind_from_stats: t = 1.8759  p = 0.120414
formula:              t = 1.8759  p = 0.120414
```



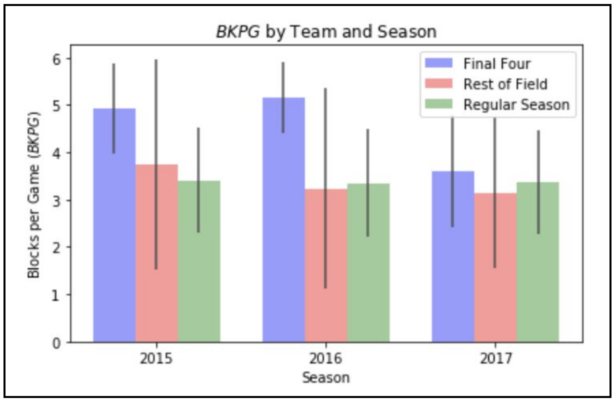


```
ttest_ind:            t = 3.67385  p = 0.0100572
ttest_ind_from_stats: t = 3.67385  p = 0.0100572
formula:              t = 3.67385  p = 0.0100572
```





```
ttest_ind:            t = 0.632449  p = 0.564705
ttest_ind_from_stats: t = 0.632449  p = 0.564705
formula:              t = 0.632449  p = 0.564705
```



The last statistic that we chose to investigate for the tournament was the average number of blocks per game a team had. Over the past three seasons, we saw that the teams that did not make the Final Four had an average number of blocks per game that came out to *3.750*, *3.236*, *3.150*. Conversely, teams that made the Final Four averaged *4.925*, *5.150*, and *3.600* blocks per game.

Using the following null hypothesis:

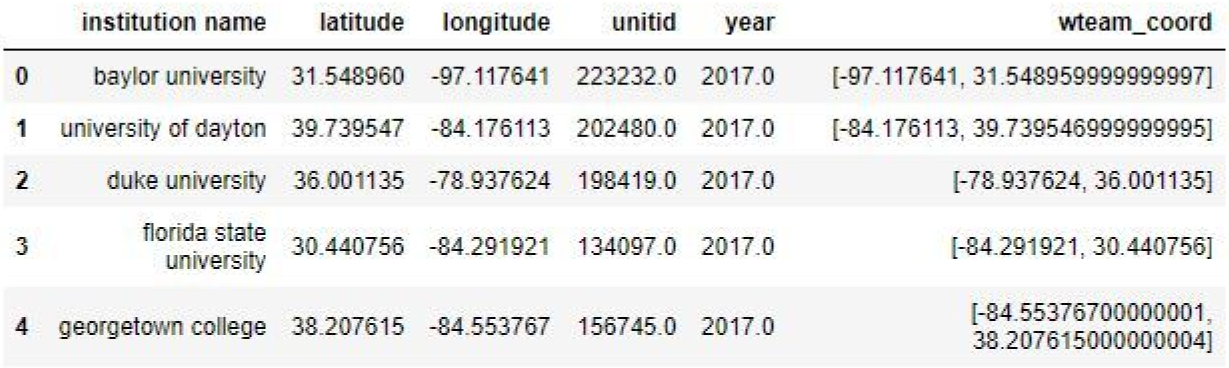
H_0*: no difference in average number of blocks per game*
H_A*: there exists a difference in the number of blocks per game*

and a significance level of *0.05* we see that our results have, on average, not shown statistical significance in this category. For the past three seasons we have *p-values* of *0.120*, *0.010*, and *0.565*. This tells us that teams that the average number of blocks per game do not play a large role in deciding whether or not a team is going to have success in the tournament and make it to the Final Four. However, we do see a similarity in our combined bar plot in that the teams that do not make the tournament possess about the same number of blocks per game as those teams that make the tournament, but fail to reach the Final Four.

From the "hard-work" statistics that we chose to identify both rebounds and fouls appear to be good indicators of how far teams will go in the NCAA tournament. Blocks and steals per game tend to happen more infrequently than the two aforementioned statistics which may be part of the reason why they do not have as much of an impact on a team's journey to the Final Four.

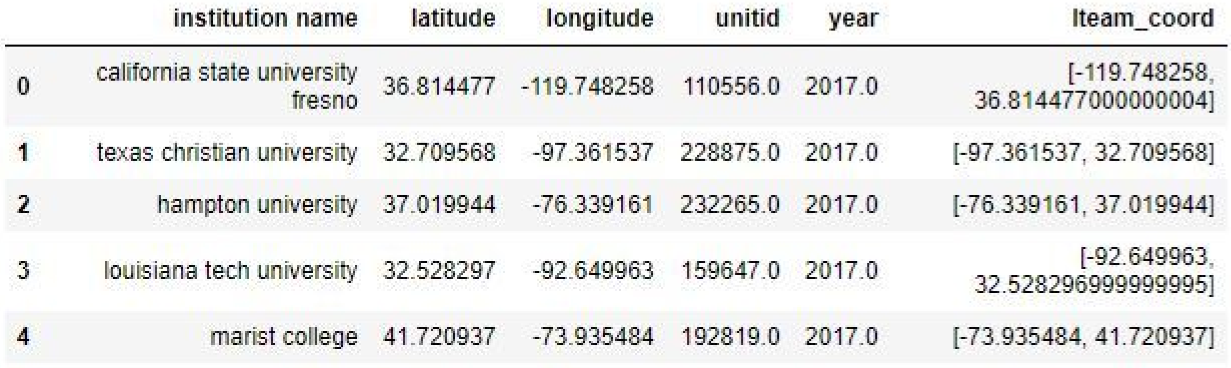
Jet-Lagged

We wanted to determine if the distance traveled to play an NCAA tournament game had any effect on a team's performance in the game, particularly if distance traveled has any effect on a team's proportion of wins. To analyze the distance that each team travelled for each game, we had to obtain additional data from the Integrated Postsecondary Education Data System, or IPEDS. Specifically, we obtained data about the institution name associated with the winning and losing teams as well as each institution's latitude and longitude coordinates, or lat-long pairs. Once we obtained the location data, we first worked with just the 2010 NCAA tournament games data since we wanted to work with only the recent data to determine the different seeds. This is known as block analysis when working with time series data. We used this kind of analysis since teams change players every few years, which changes each team's win percentage. Here are the first five winning teams and each of their respective institution coordinates for the 2010 NCAA tournament:



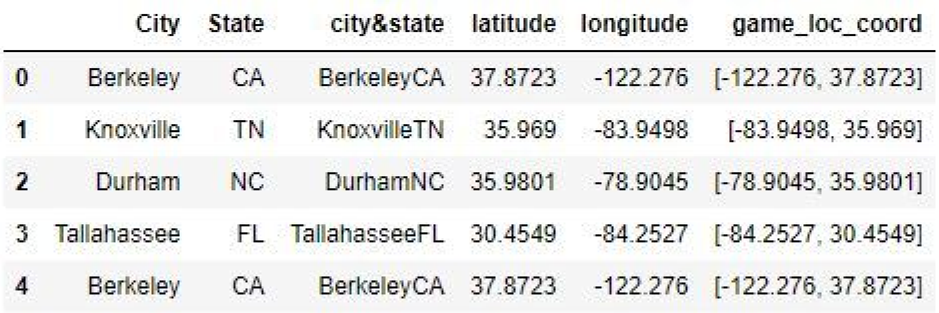
	institution name	latitude	longitude	unitid	year	wteam_coord
0	baylor university	31.548960	-97.117641	223232.0	2017.0	[-97.117641, 31.548959999999997]
1	university of dayton	39.739547	-84.176113	202480.0	2017.0	[-84.176113, 39.739546999999995]
2	duke university	36.001135	-78.937624	198419.0	2017.0	[-78.937624, 36.001135]
3	florida state university	30.440756	-84.291921	134097.0	2017.0	[-84.291921, 30.440756]
4	georgetown college	38.207615	-84.553767	156745.0	2017.0	[-84.55376700000001, 38.207615000000004]

Here are the first five losing teams and each of their respective institution coordinates for the 2010 NCAA tournament:



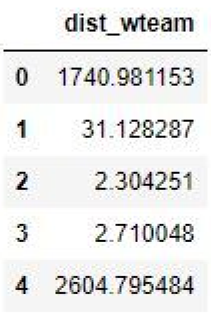
	institution name	latitude	longitude	unitid	year	lteam_coord
0	california state university fresno	36.814477	-119.748258	110556.0	2017.0	[-119.748258, 36.814477000000004]
1	texas christian university	32.709568	-97.361537	228875.0	2017.0	[-97.361537, 32.709568]
2	hampton university	37.019944	-76.339161	232265.0	2017.0	[-76.339161, 37.019944]
3	louisiana tech university	32.528297	-92.649963	159647.0	2017.0	[-92.649963, 32.528296999999995]
4	marist college	41.720937	-73.935484	192819.0	2017.0	[-73.935484, 41.720937]

Here are the first five game locations and each of their respective location coordinates for the 2010 NCAA tournament:

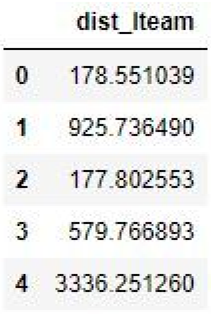


	City	State	city&state	latitude	longitude	game_loc_coord
0	Berkeley	CA	BerkeleyCA	37.8723	-122.276	[-122.276, 37.8723]
1	Knoxville	TN	KnoxvilleTN	35.969	-83.9498	[-83.9498, 35.969]
2	Durham	NC	DurhamNC	35.9801	-78.9045	[-78.9045, 35.9801]
3	Tallahassee	FL	TallahasseeFL	30.4549	-84.2527	[-84.2527, 30.4549]
4	Berkeley	CA	BerkeleyCA	37.8723	-122.276	[-122.276, 37.8723]

Once we got the location coordinates for each winning team, losing team, and game location, we calculated the distance traveled for each winning and losing team to their respective game location using the haversine distance formula, which calculates the shortest distance between two points on a sphere, which in this case is Earth, using their latitude and longitude coordinates measured along the surface. Here are the haversine distances for the first five winning and losing teams, respectively:

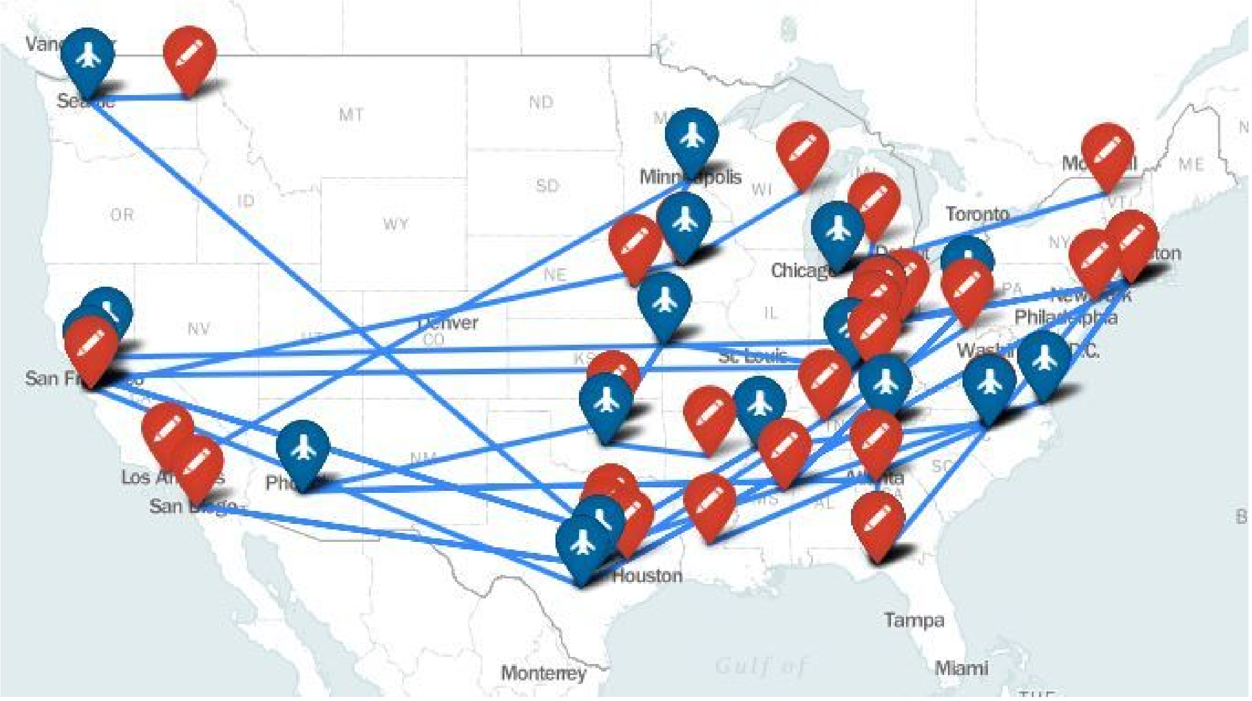


	dist_wteam
0	1740.981153
1	31.128287
2	2.304251
3	2.710048
4	2604.795484

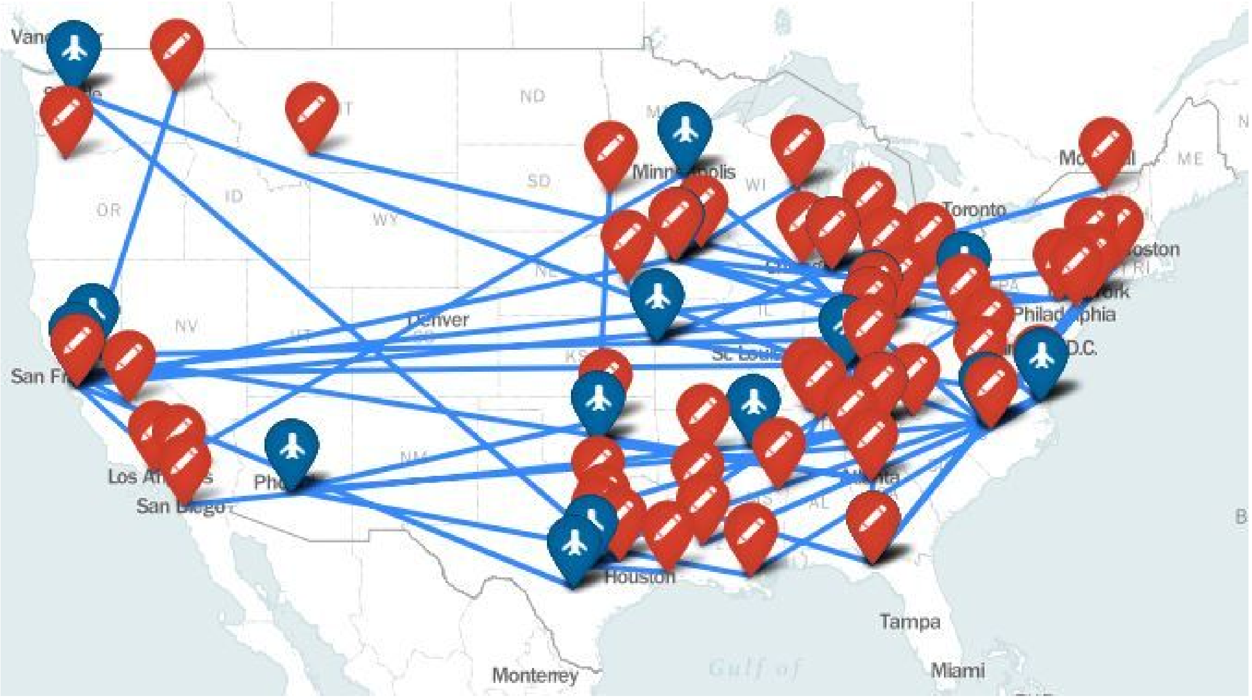


	dist_lteam
0	178.551039
1	925.736490
2	177.802553
3	579.766893
4	3336.251260

Then, we created GeoJSON files to plot distances traveled for winning teams and losing teams. For these maps, red icons with pencils indicate an institution, and blue icons with planes indicate a game location. Here is the map showing distances traveled for winning teams:

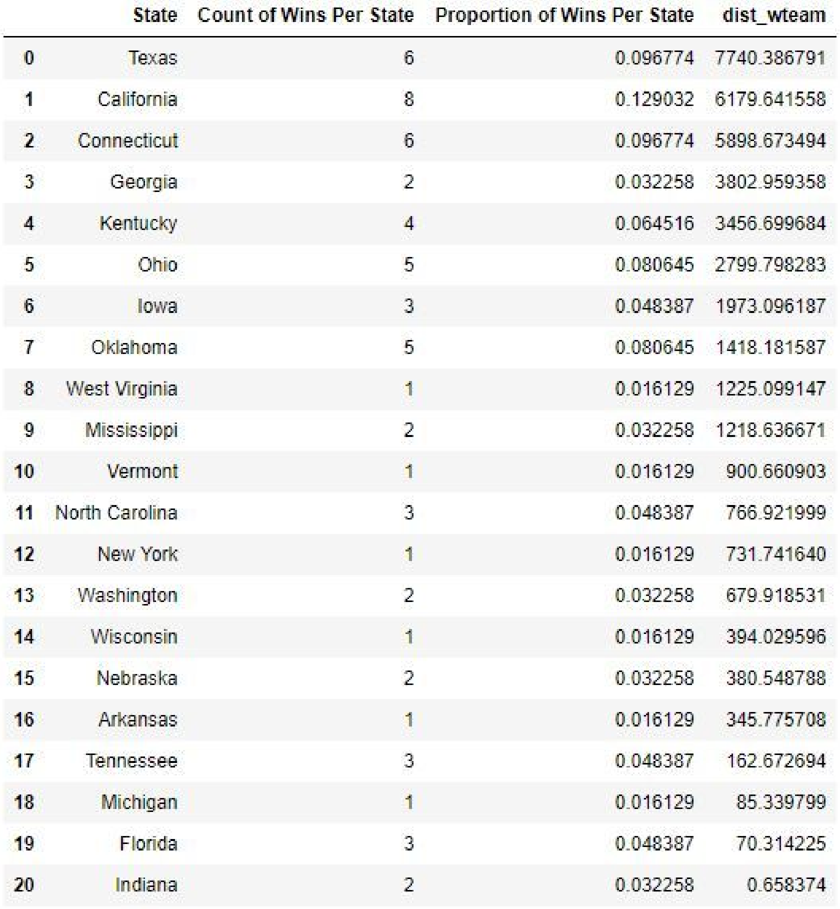


Here is the map showing distances traveled for losing teams:

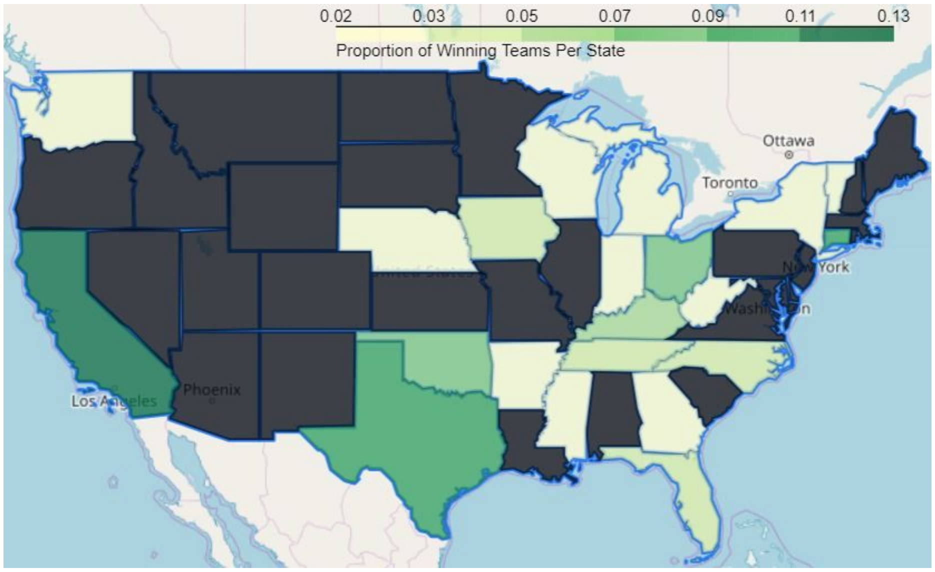


As we can see from the maps, a majority of institutions and game locations are from the Midwest to East Coast, with only a few institutions and game location being from the West Coast. Since we are determining if distance traveled has any effect on proportion of wins, we then calculated both the counts and proportion of wins per state to first see if there is any relationship between distance traveled and proportion of wins for each state. In addition, we also calculated the total distance traveled for the winning teams for each state to see if there is a relationship between distance traveled and proportion of wins.

Here are the counts and proportions of wins per state as well as distance traveled for winning teams per state and the corresponding choropleth map:

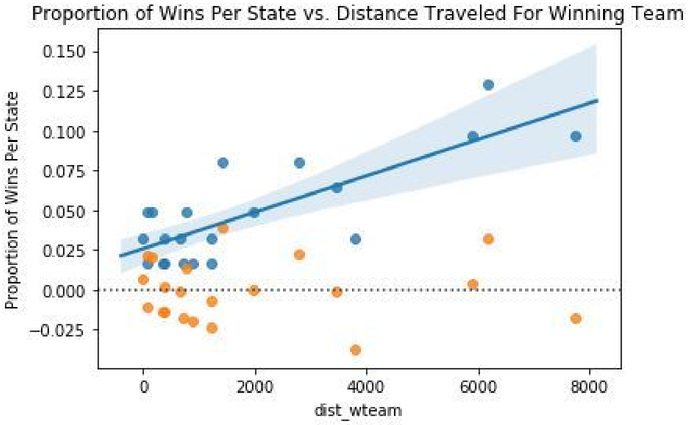


	State	Count of Wins Per State	Proportion of Wins Per State	dist_wteam
0	Texas	6	0.096774	7740.386791
1	California	8	0.129032	6179.641558
2	Connecticut	6	0.096774	5898.673494
3	Georgia	2	0.032258	3802.959358
4	Kentucky	4	0.064516	3456.699684
5	Ohio	5	0.080645	2799.798283
6	Iowa	3	0.048387	1973.096187
7	Oklahoma	5	0.080645	1418.181587
8	West Virginia	1	0.016129	1225.099147
9	Mississippi	2	0.032258	1218.636671
10	Vermont	1	0.016129	900.660903
11	North Carolina	3	0.048387	766.921999
12	New York	1	0.016129	731.741640
13	Washington	2	0.032258	679.918531
14	Wisconsin	1	0.016129	394.029596
15	Nebraska	2	0.032258	380.548788
16	Arkansas	1	0.016129	345.775708
17	Tennessee	3	0.048387	162.672694
18	Michigan	1	0.016129	85.339799
19	Florida	3	0.048387	70.314225
20	Indiana	2	0.032258	0.658374



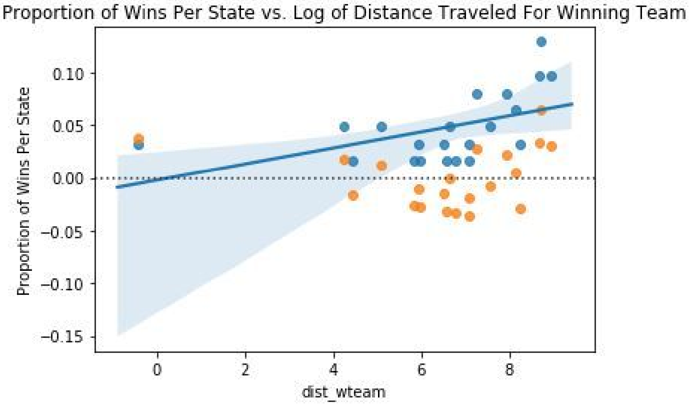
To clarify, we sorted by distance traveled to find if there is a relationship. As we can see from the dataframe above, it looks like distance traveled does not follow a clear relationship with proportion of wins since there are some inconsistencies with these features. For example, Georgia had 2 wins out of 63 total games, which is about a 0.0322 proportion of wins or about a 3.22% win rate, and traveled about 3,802 miles to play NCAA tournament games. For comparison, Ohio had 5 wins out of 63 total games, which is about a 0.08 proportion of wins or about an 8% win rate, and traveled only about 2,799 miles, which means that winning teams from Ohio traveled about 1,003 miles less than winning teams from Georgia, yet the Ohio teams won 3 more games than Georgia teams. A similar insight is seen West Virginia, with only 1 win and about 1,225 miles traveled, and North Carolina, with 3 wins and about 766 miles traveled. Specifically, North Carolina teams traveled about 459 miles less than West Virginia teams, yet the North Carolina teams won 2 more games than West Virginia teams.

Now that we have seen, from quantitative data, some inconsistencies with the data that seems to refute the idea that there is a relationship between distance traveled and proportion of wins, let's see if we can determine a relationship using the concept of linear regression and the least squares line, which is detailed in the theory portion. Here is the linear regression plot in blue and the residuals plot in orange:



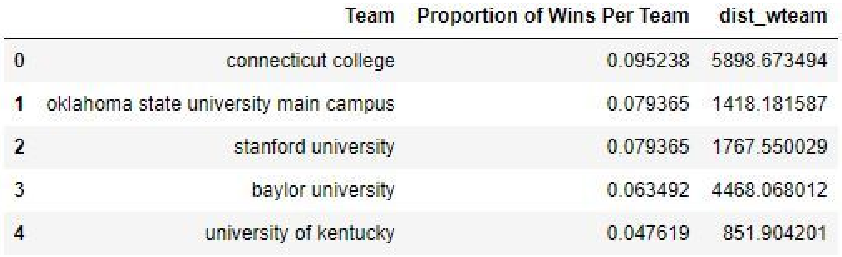
As we can see from the above plot, the data does not look linear, so a linear regression model would not be appropriate. When performing the linear regression, we found that the r-squared value is 0.632031, indicating a non-ideal fit. Additionally, we do not see nearly normal residuals and inconstant variability, all of which do not meet the conditions for the least squares line.

From here, we considered a variety of transformations of the data, such that the linear regression model would be a more appropriate fit. We tried a logarithmic transformation of the proportion of wins to see if this approach would yield the most linear scatterplot as well as small and reasonably normal residuals. I applied a log transformation to the distance traveled in the plot below, where the linear regression plot is in blue and the residuals plot is in orange:



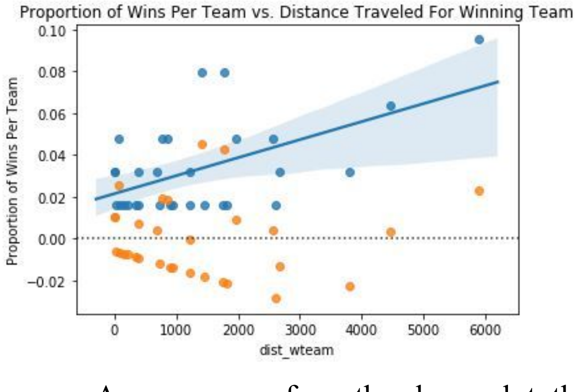
Again, we do not see any linearity between distance traveled and proportion of wins. In addition, we do not see nearly normal residuals and inconstant variability, all of which do not meet the conditions for the least squares line. Therefore, we cannot legitimately perform linear regression.

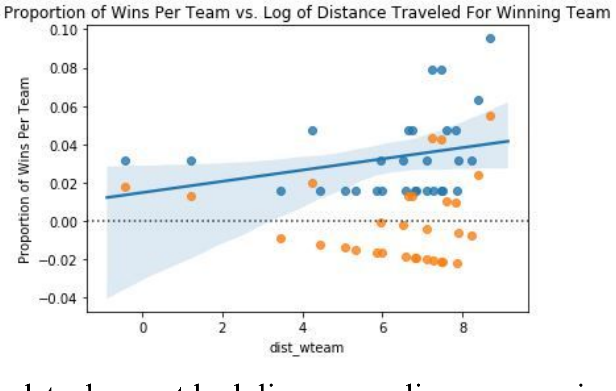
Then, we tried to apply the same analysis to each team rather than each state. Here is a dataframe showing the five institutions with teams that traveled the most and each of their respective proportion of wins:



	Team	Proportion of Wins Per Team	dist_wteam
0	connecticut college	0.095238	5898.673494
1	oklahoma state university main campus	0.079365	1418.181587
2	stanford university	0.079365	1767.550029
3	baylor university	0.063492	4468.068012
4	university of kentucky	0.047619	851.904201

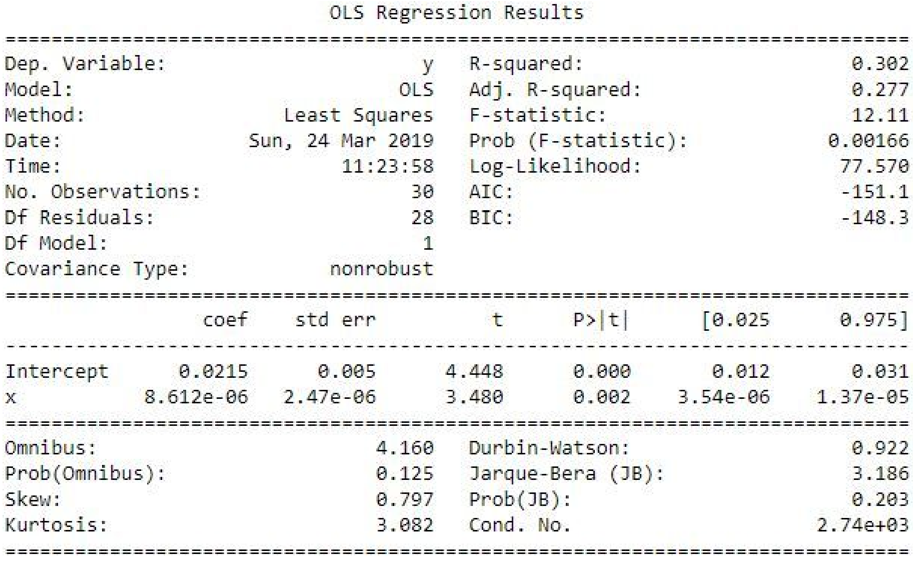
On the left is the linear regression plot in blue and the residuals plot in orange. On the right, we applied a log transformation to the distance traveled in the plot below, where the linear regression plot is in blue and the residuals plot is in orange.





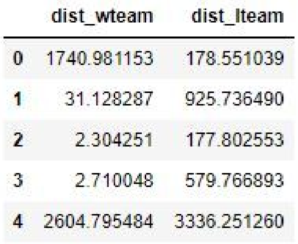
As we can see from the above plot, the data does not look linear, so a linear regression model would not be appropriate. When performing the linear regression, we found that the r-squared value is 0.301981, indicating a non-ideal fit, worse so than the data per state. From here, we considered a variety of transformations of the data, such that the linear regression model would be a more appropriate fit. We tried a logarithmic transformation of the proportion of wins to see if this approach would yield the most linear scatterplot as well as small and reasonably normal residuals. Again, we do not see any linearity between distance traveled and proportion of wins. In addition, we do not see nearly normal residuals and inconstant variability, all of which do not meet the conditions for the least squares line. Therefore, we cannot legitimately perform linear regression.

Here are the results when performing ordinary least squares regression, for reference:

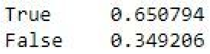


```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.302
Model:                            OLS   Adj. R-squared:                  0.277
Method:                 Least Squares   F-statistic:                     12.11
Date:                Sun, 24 Mar 2019   Prob (F-statistic):            0.00166
Time:                        11:23:58   Log-Likelihood:                 77.570
No. Observations:                  30   AIC:                            -151.1
Df Residuals:                      28   BIC:                            -148.3
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0215      0.005      4.448      0.000       0.012       0.031
x           8.612e-06   2.47e-06      3.480      0.002    3.54e-06    1.37e-05
==============================================================================
Omnibus:                        4.160   Durbin-Watson:                   0.922
Prob(Omnibus):                  0.125   Jarque-Bera (JB):                3.186
Skew:                           0.797   Prob(JB):                        0.203
Kurtosis:                       3.082   Cond. No.                     2.74e+03
==============================================================================
```

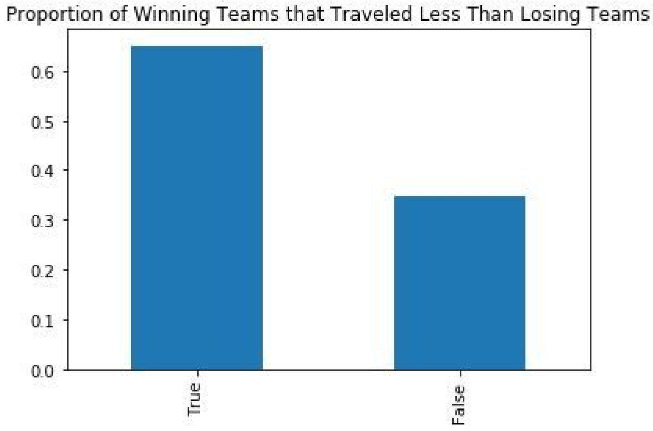
Although we cannot perform linear regression, we still wanted to see if there is any relationship between distance traveled and proportion of wins. So, we determined the proportions of winning teams that traveled less than losing teams and plotted the proportions as follows:



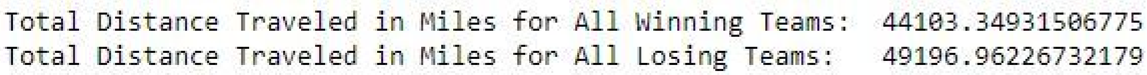
	dist_wteam	dist_lteam
0	1740.981153	178.551039
1	31.128287	925.736490
2	2.304251	177.802553
3	2.710048	579.766893
4	2604.795484	3336.251260



```
True     0.650794
False    0.349206
```



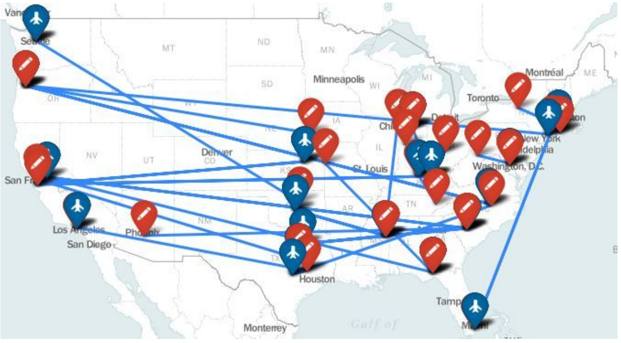
Proportion of Winning Teams that Traveled Less Than Losing Teams

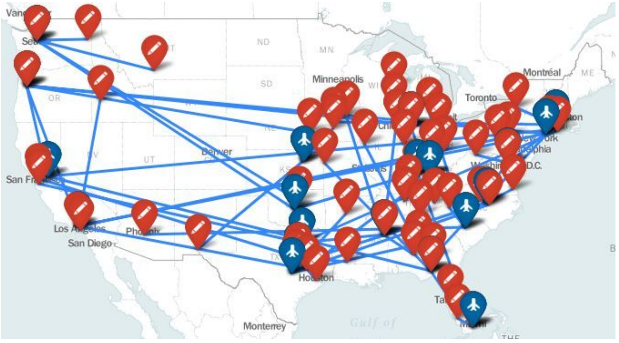


```
Total Distance Traveled in Miles for All Winning Teams:  44103.34931506775
Total Distance Traveled in Miles for All Losing Teams:   49196.96226732179
```

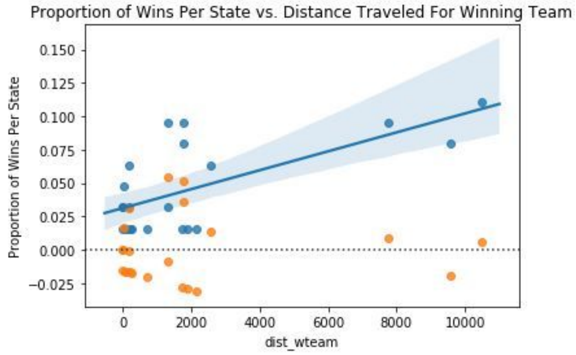
As we can see from the quantitative data and accompanying plot, we see that a larger proportion of winning teams travel less than losing teams. Specifically, about 65% of winning teams travel less than losing teams while only about 34% of winning teams actually travel more than losing teams. So, although we did not find a clear linear relationship between distance traveled and proportion of wins, we do find that the event of the winning team traveling less than the losing team occurs more frequently than the event of the winning team traveling more than the losing team. In addition, we see that, when the distance traveled is combined into all winning teams and all losing teams, we see that the total distance traveled for all losing teams is about 5,000 miles more than the total distance traveled for all winning teams. This insight further implies the notion that there is some connection between distance traveled and proportion of wins.

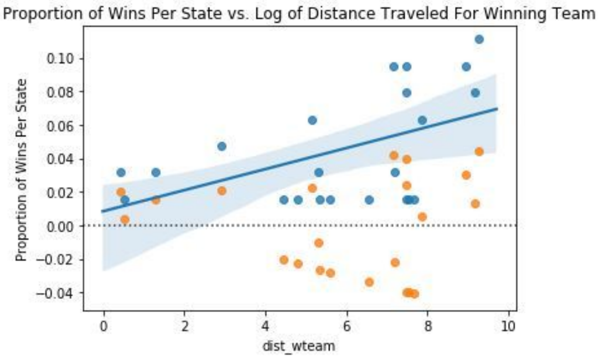
To determine whether or not this insight is just a coincidence based off of data from only one year, we decided to perform the same analysis on the 2017 NCAA tournament season. By using one of the more recent seasons, we can determine if there is any continuity or change in the data over time, due to confounding factors, such as that teams change players every few years, which changes each team's win percentage. So, we matched each winning team, losing team, and game location to their respective location coordinates. Then, we plotted maps showing distances traveled for winning teams and losing teams, respectively:





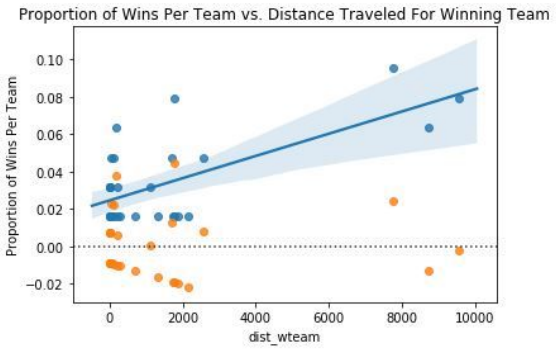
As we can see from the maps, there is a similarity to the 2010 season in that a majority of institutions and game locations are from the Midwest to East Coast, with only a few institutions and game location being from the West Coast. Next, we performed linear regression on the 2017 season data to determine if there is a relationship between distance traveled and proportion of wins that did not occur in the 2010 season. Here are the plots for, per state, proportion of wins versus distance traveled, and the logarithmic transformation of distance traveled, respectively:

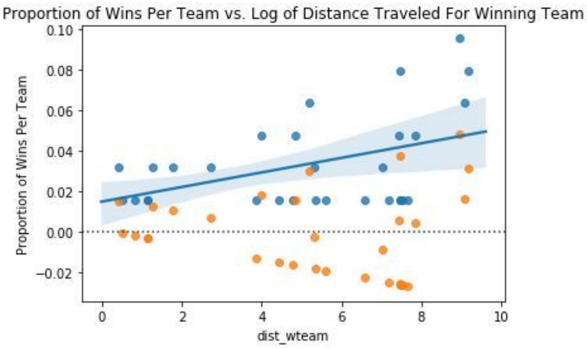




As we can see from the above plot, the data does not look linear, so a linear regression model would not be appropriate. When performing the linear regression, we computed an r-squared value of 0.433938, which indicates a non-ideal fit. Additionally, we do not see nearly normal residuals and inconstant variability, all of which do not meet the conditions for the least squares line.

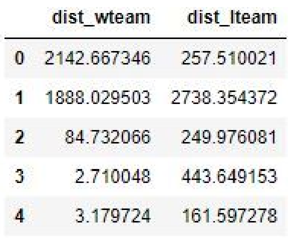
Here are the plots for, per team, proportion of wins versus distance traveled, and the logarithmic transformation of distance traveled, respectively:


Proportion of Wins Per Team vs. Distance Traveled For Winning Team

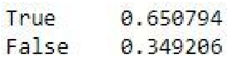

Proportion of Wins Per Team vs. Log of Distance Traveled For Winning Team

As we can see from the above plot, the data does not look linear, so a linear regression model would not be appropriate. When performing the linear regression, we computed an r-squared value of 0.447937, which also indicates a non-ideal fit. Additionally, we do not see nearly normal residuals and inconstant variability, all of which do not meet the conditions for the least squares line. Therefore, we could not legitimately perform linear regression.

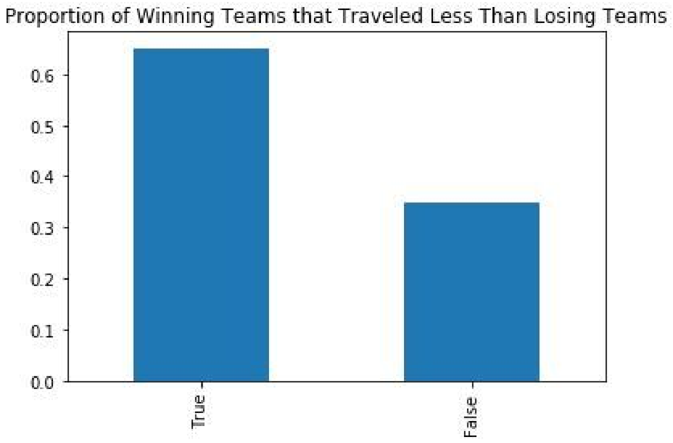
Here are the proportions of winning teams that traveled less than losing teams and plot of the proportions:

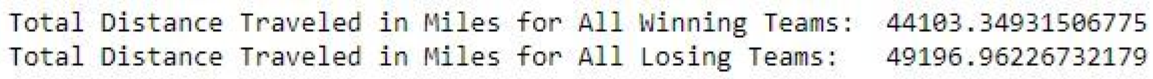


	dist_wteam	dist_lteam
0	2142.667346	257.510021
1	1888.029503	2738.354372
2	84.732066	249.976081
3	2.710048	443.649153
4	3.179724	161.597278



```
True     0.650794
False    0.349206
```

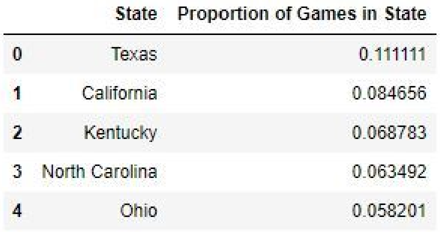

Proportion of Winning Teams that Traveled Less Than Losing Teams



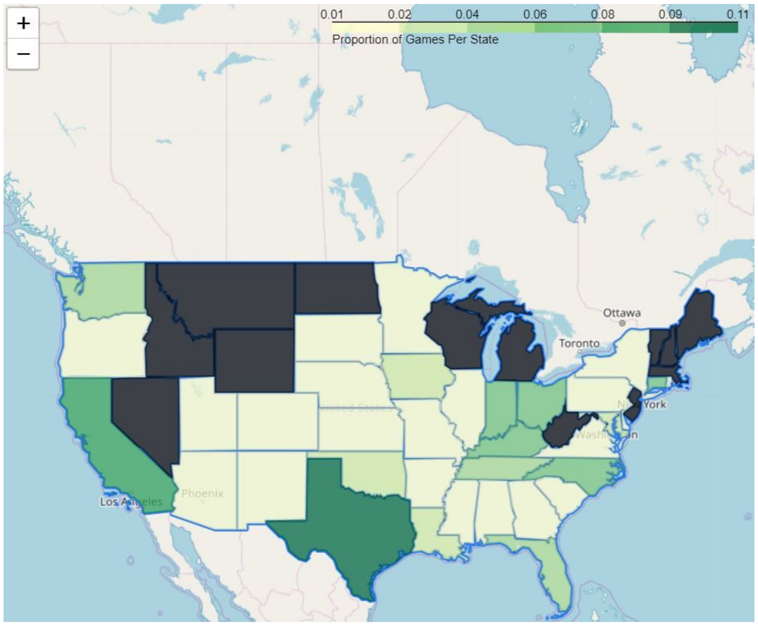
```
Total Distance Traveled in Miles for All Winning Teams:  44103.34931506775
Total Distance Traveled in Miles for All Losing Teams:   49196.96226732179
```

As we can see from the quantitative data and accompanying plot, we see that a larger proportion of winning teams travel less than losing teams. Specifically, about 65% of winning teams travel less than losing teams while only about 34% of winning teams actually travel more than losing teams. Surprisingly, although each of the winning and losing teams traveled distances that are different to distance traveled from the teams playing during the 2010 season, the total distances traveled for all winning teams and for all losing teams for the 2017 season were identical to those for the 2010 season. In addition, the proportion of winning teams that traveled less than losing teams for the 2017 season was also identical to that for the 2010 season, with about 65% of winning teams traveling less than losing teams. Even though that this occurrence could simply be due to a coincidence, this insight implies that throughout time, winning teams tend to travel less than losing teams. Therefore, throughout time, there seems to be this implication that a majority of the time, if team A travels a shorter distance than team B, then team A will have a better chance of winning the game. This insight makes sense since all teams are subject to some effect of long distance travel (especially by plane), such as jet lag, grogginess, or some other adverse effect. In the example, if team A travels a shorter distance than team B, then team A will experience less adverse effects, such as jet lag, allowing team A's players to potentially perform better than team B.

Additionally, we calculated the proportion of all games, starting with the 2010 season and from both the regular season and the NCAA tournament, played per state and plotted the proportion of all games on a choropleth map, both of which are displayed here:



	State	Proportion of Games in State
0	Texas	0.111111
1	California	0.084656
2	Kentucky	0.068783
3	North Carolina	0.063492
4	Ohio	0.058201

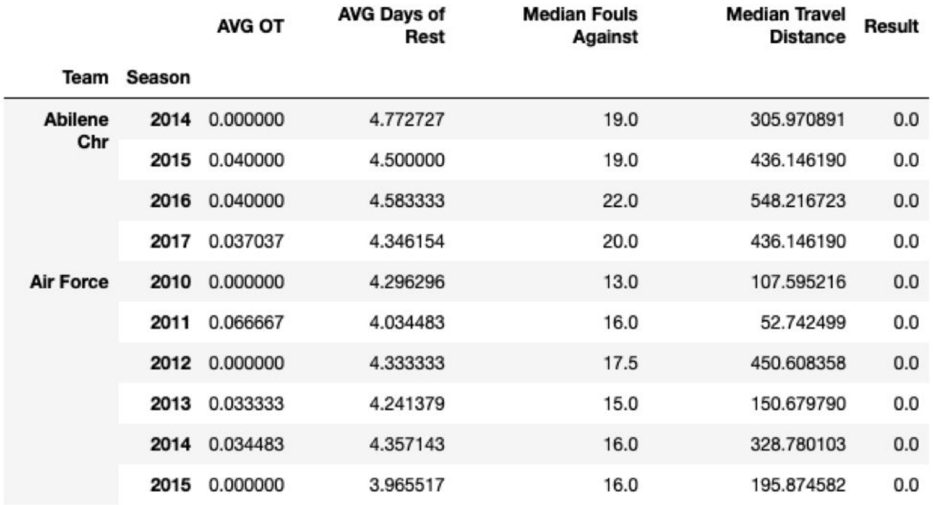


As we can see from the choropleth map, the state with the highest proportion of games is Texas, which is where about 11.1% of games are played out of all games, starting with the 2010 season. After Texas, we find that California has about 8% of games played, Kentucky has about 6.8% of games played, North Carolina has about 6.3% of games played, and Ohio has about 5.8% of games played. So, we see that 4 out of the top 5 states are states located in the Midwest to East Coast regions, with the only exception being California.

Practice vs. Wear

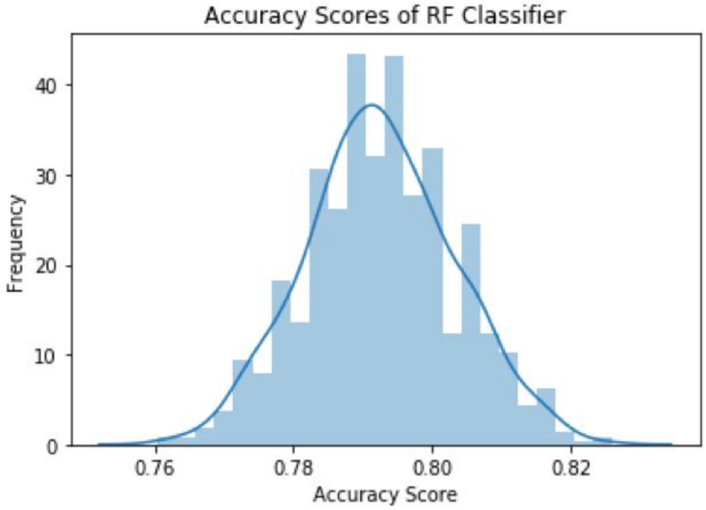
In developing our final predictive model, we sought to create a feature that quantifies how “beat-up” a team is at the end of their season. Our group reasoned that this “beat-up” metric should take into account how often a team went into overtime, how much rest they received between games, the number of times they were fouled a game and the distance they had to travel to their games. Each of these factors takes a toll on a team’s players over the course of a season and should be addressed in predicting performance. In creating this metric, our team decided to use the average number of games that went into overtime, the average days of rest in between games, the median number of fouls committed against a team’s players, and the median distance traveled by a team over the course of a season to quantify the above factors. We hypothesized that there would be a negative correlation between this beat-up metric and a team’s success, and that high values of this beat-up feature would help our final model identify those teams that didn’t make the NCAA tournament.

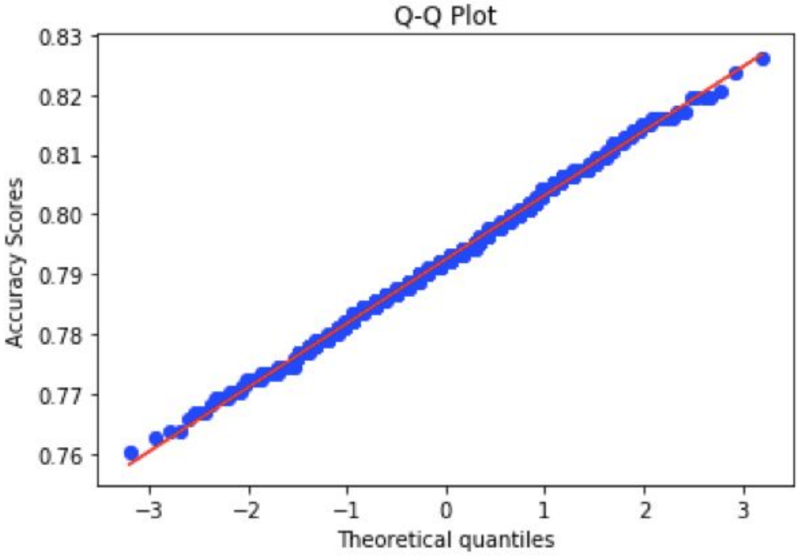
To begin engineering this feature, we created a table composed of the average number of games that went into overtime, the average days of rest in between games, the median number of fouls committed against a team’s players and the median distance traveled the team that competed in each of the seasons from 2010-2017. Note that “Median Travel Distance” variable is in miles and takes into account the total distance traveled to a game and back. As geographical data was not included in the provided datasets, longitude and latitude data was imported from an external source. Once longitude and latitude data was added, distances between locations was calculated using the haversine formula. In addition to these metrics, a “Result” column was added that quantified a team’s success: 0 indicates a team didn’t make the NCAA playoffs, 1 indicates a team made the NCAA playoffs but didn’t make the round of 16, 2 indicated a team made the round of 16 but didn’t make the final 4 and 3 indicated a team made the final 4. The first 10 entries of this table are displayed below:



		AVG OT	AVG Days of Rest	Median Fouls Against	Median Travel Distance	Result
Team	Season					
Abilene Chr	2014	0.000000	4.772727	19.0	305.970891	0.0
	2015	0.040000	4.500000	19.0	436.146190	0.0
	2016	0.040000	4.583333	22.0	548.216723	0.0
	2017	0.037037	4.346154	20.0	436.146190	0.0
Air Force	2010	0.000000	4.296296	13.0	107.595216	0.0
	2011	0.066667	4.034483	16.0	52.742499	0.0
	2012	0.000000	4.333333	17.5	450.608358	0.0
	2013	0.033333	4.241379	15.0	150.679790	0.0
	2014	0.034483	4.357143	16.0	328.780103	0.0
	2015	0.000000	3.965517	16.0	195.874582	0.0

To determine if the above variables are useful in predicting a team's success, a Random Forest Classifier was ran using Scikit-Learn. A 95% confidence interval for the true accuracy score of the predictor was found to be (0.792,0.793). The interval was created by splitting our data into training and testing data 1,000 times and calculating the predictor's accuracy score for each iteration. The confidence interval was then computed by taking the interval created by adding and subtracting 1.96*(the standard error of the means) from the mean of the accuracy scores. To verify the normality of the distribution of accuracy scores, a histogram with a KDE overlay and a qq-plot are shown below:

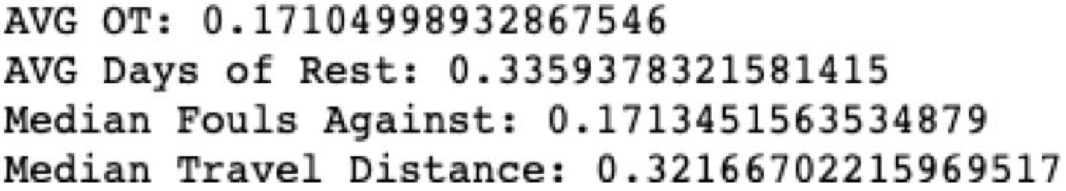




Since the histogram appear uniform and the data points in the qq plot deviate very little from the red straight line, the distribution of accuracy scores appear normal. Thus the 95% confidence interval calculated is legitimate. This confidence interval implies that we can be 95%

certain that this interval contains the true accuracy score of the random forest classifier. Our calculated accuracy score of 79% is roughly the same as a constant model that predicts whether a team won't make the NCAA tournament since roughly 80% of the teams don't make the NCAA tournament. However, this engineered "beat-up" metric will be supplemented with other features in the final model, thus this model's accuracy score isn't important alone .

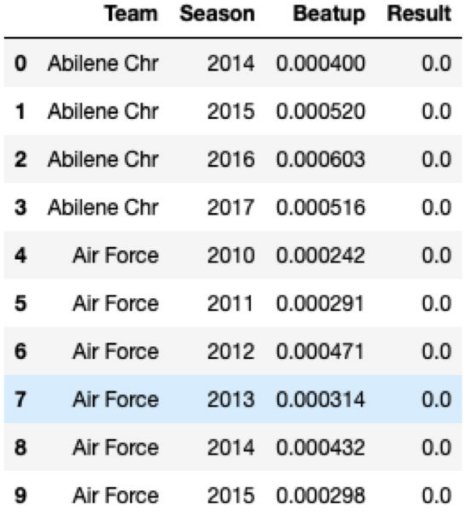
To create this beat-up metric, the proportion each of the above features contributed to predicting team success (i.e, result of 0,1,2 or 3), was found using a random forest classifier. Using scikit-learn's feature importance method, the averages of the importance of each feature in 1000 random forest classifier experiments is shown below:



```
AVG OT: 0.17104998932867546
AVG Days of Rest: 0.3359378321581415
Median Fouls Against: 0.1713451563534879
Median Travel Distance: 0.32166702215969517
```

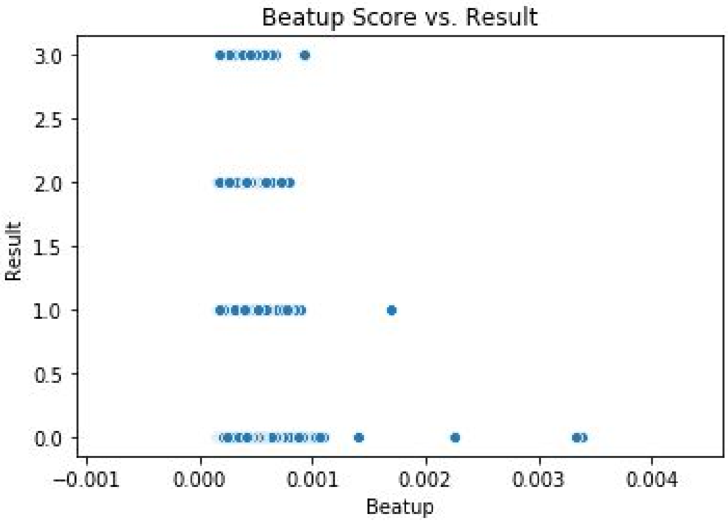
To make the scale of the variables similar, the AVG OT, AVG Days of Rest, Median Fouls Against and Median Travel Distance columns were normalized. To aggregate these normalized values into a team's beat-up score, for each team, the values of each of the above features were multiplied by the feature's corresponding "importance" score. Doing so weighted each of the features by how important they were in predicting a team's success; a variable that had more predictive power will be given more weight in the resulting beat-up metric than one with less predictive power. The beat-up metric was then created by summing weighted values.

Once the beat-up metric was calculated for each team in each season, a final table was produced containing a team's name, the season, the beat-up score and the result of their season. The first 10 entries of this table are displayed below:

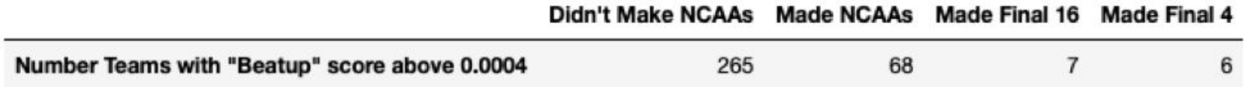


	Team	Season	Beatup	Result
0	Abilene Chr	2014	0.000400	0.0
1	Abilene Chr	2015	0.000520	0.0
2	Abilene Chr	2016	0.000603	0.0
3	Abilene Chr	2017	0.000516	0.0
4	Air Force	2010	0.000242	0.0
5	Air Force	2011	0.000291	0.0
6	Air Force	2012	0.000471	0.0
7	Air Force	2013	0.000314	0.0
8	Air Force	2014	0.000432	0.0
9	Air Force	2015	0.000298	0.0

A scatter plot with the beat-up variable as the explanatory variable and the Result variable as the response variable is shown below:

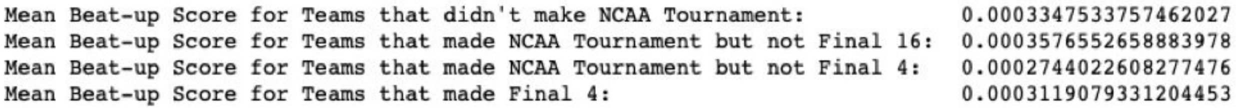


While there is much overlap in the beat-up scores among the different values of the Result variable, it appears that teams that didn't make the NCAA tournament are more likely to have higher beat-up scores. This is also shown by the counts of schools that had a beat-up score greater than 0.005:



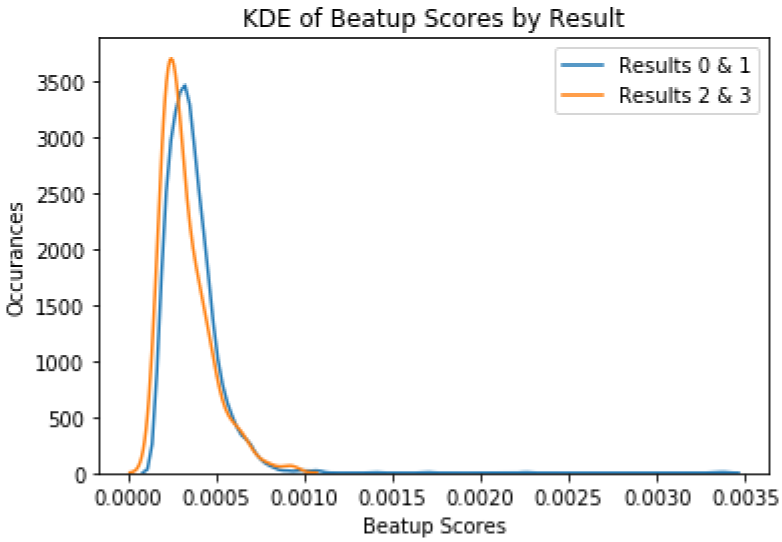
	Didn't Make NCAAs	Made NCAAs	Made Final 16	Made Final 4
Number Teams with "Beatup" score above 0.0004	265	68	7	6

Furthermore, the mean beat-up scores for teams in each of the results categories are shown below:



```
Mean Beat-up Score for Teams that didn't make NCAA Tournament:          0.0003347533757462027
Mean Beat-up Score for Teams that made NCAA Tournament but not Final 16: 0.0003576552658883978
Mean Beat-up Score for Teams that made NCAA Tournament but not Final 4:  0.0002744022608277476
Mean Beat-up Score for Teams that made Final 4:                          0.0003119079331204453
```

The mean scores of teams that made the final 16 or final 4 are lower than the mean scores of teams that did not make the final 16 or final 4. This further indicates our beat-up metric as a viable feature to add to our final predictive model. However, KDE plots of beat-up scores for teams that didn't make it as far as the NCAA tournament round of 16 and those that did are produced, the distributions of beat-up scores conditioned on the Result variable appear similar:



To determine if there is a statistical difference in the means of the beat-up metric when conditioned on results, we will conduct a permutation test at a 0.05 significance level. For simplicity let us abbreviate the absolute value of the difference of means as ADOM. For this test, the null hypothesis is that the ADOM between the two groups is 0. Thus, the alternative hypothesis is that the ADOM between the two groups is not zero. The observed ADOM between the two groups (one group is composed of teams with result 0 or 1 and the other group is composed of teams with result 2 or 3) is 0.000035. The permutation test was performed by:

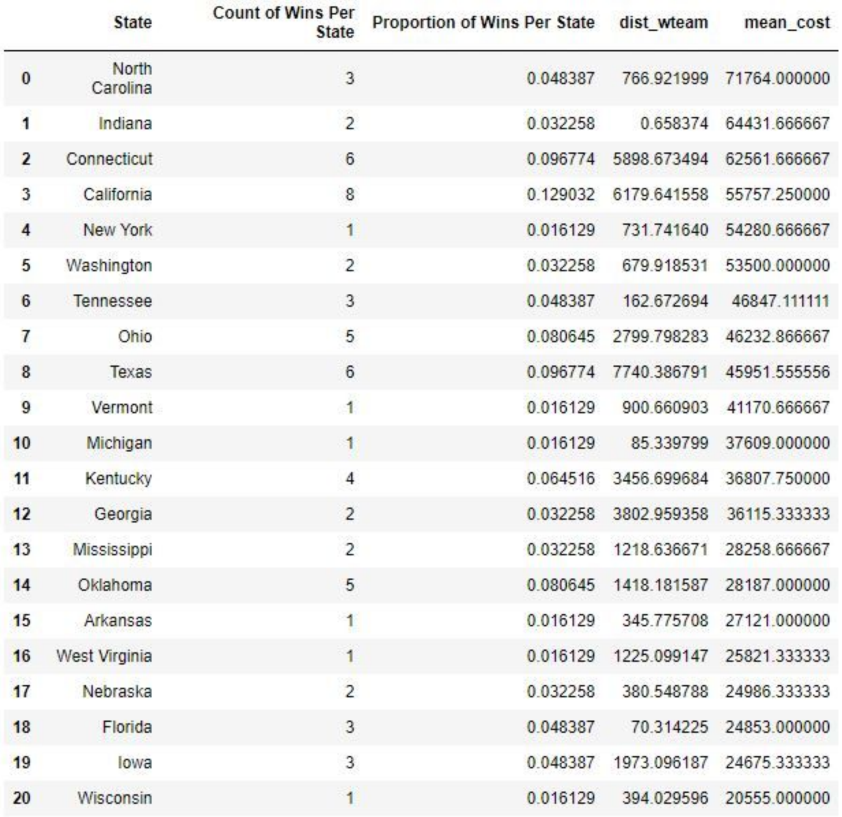
1) shuffling the beat-up scores so that they were randomly assigned to the results category
2) grouped the shuffled table by the result groups
3) calculate ADOM of these two grouped tables

This test was performed 10,000 times, each time calculating the ADOM. After doing so, the p-value of the test was found to be 0.0213. This means that under the null hypothesis, the probability of seeing an ADOM at least as extreme as the observed one is 0.0229. Thus with a 0.05 significance level, we reject the null hypothesis and accept the alternative hypothesis.

Since it was shown that there exists a statistical difference in the means of the beat-up scores between teams that made the final 16 of the NCAA tournament and those that did not, the beat-up metric may a valuable feature to utilize in our final model.

Additional Inquiries

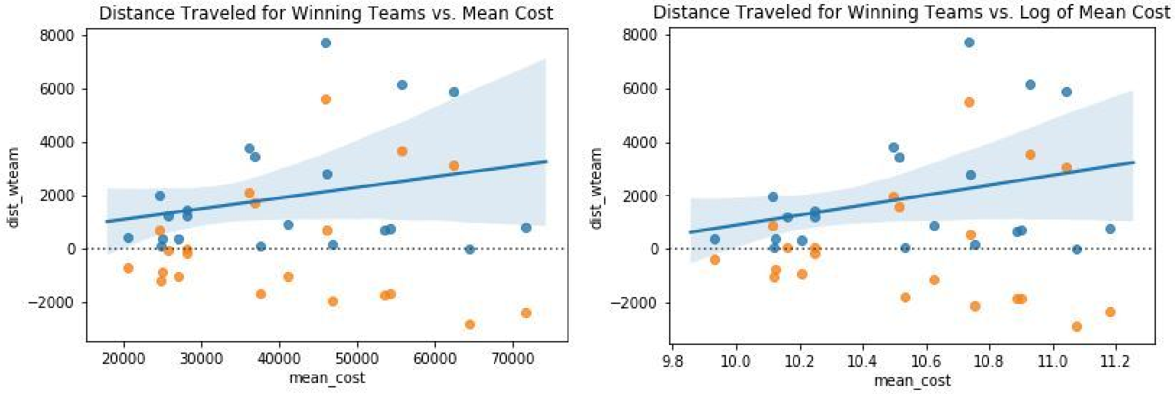
For our additional hypothesis, we wanted to see if there is any relationship between the average cost of attendance per state and either total travel distance per state or each state's proportion of wins. We decided to analyze these specific features because we had this assumption that students who attend institutions with higher costs of attendance have the financial stability to pay for not only school expenses, but also extracurricular expenses, such as sports expenses, in this case. Specifically, we assumed that these students have the ability to pay for more access to better facilities, equipment, and amenities, resulting in better student athletes and a better chance to win games. To determine if a relationship existed between these features, we obtained additional data from IPEDS about the cost of attendance for in-state, out-of-state not with family, and out-of-state without family. Then, we took the average cost of attendance per state to create some standard measure of cost of attendance. For now, we will work with just the 2010 season data. Here are the average cost of attendance per state:



	State	Count of Wins Per State	Proportion of Wins Per State	dist_wteam	mean_cost
0	North Carolina	3	0.048387	766.921999	71764.000000
1	Indiana	2	0.032258	0.658374	64431.666667
2	Connecticut	6	0.096774	5898.673494	62561.666667
3	California	8	0.129032	6179.641558	55757.250000
4	New York	1	0.016129	731.741640	54280.666667
5	Washington	2	0.032258	679.918531	53500.000000
6	Tennessee	3	0.048387	162.672694	46847.111111
7	Ohio	5	0.080645	2799.798283	46232.866667
8	Texas	6	0.096774	7740.386791	45951.555556
9	Vermont	1	0.016129	900.660903	41170.666667
10	Michigan	1	0.016129	85.339799	37609.000000
11	Kentucky	4	0.064516	3456.699684	36807.750000
12	Georgia	2	0.032258	3802.959358	36115.333333
13	Mississippi	2	0.032258	1218.636671	28258.666667
14	Oklahoma	5	0.080645	1418.181587	28187.000000
15	Arkansas	1	0.016129	345.775708	27121.000000
16	West Virginia	1	0.016129	1225.099147	25821.333333
17	Nebraska	2	0.032258	380.548788	24986.333333
18	Florida	3	0.048387	70.314225	24853.000000
19	Iowa	3	0.048387	1973.096187	24675.333333
20	Wisconsin	1	0.016129	394.029596	20555.000000

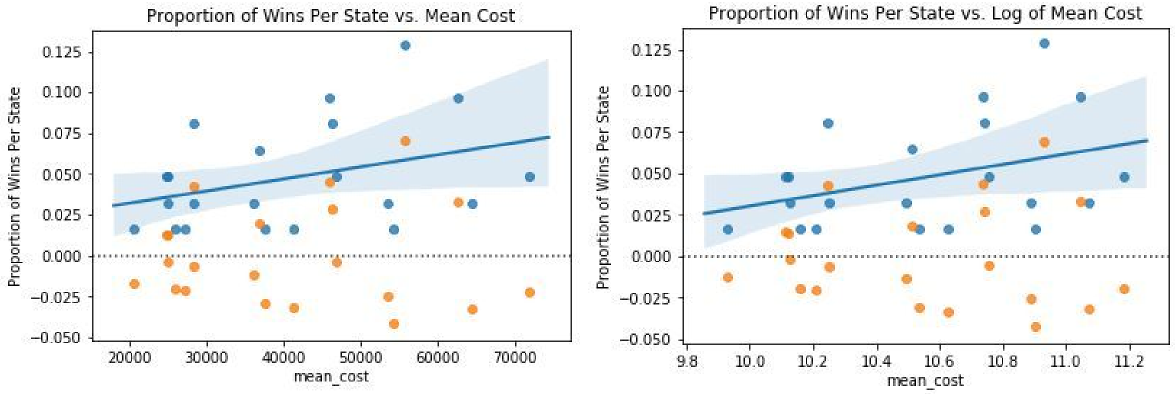
To clarify, we sorted by mean cost to find if there is a relationship. As we can see from the dataframe above, it looks like mean cost does not follow a clear relationship with proportion of wins and travel distance since there are some inconsistencies with these features. For example, Oklahoma had 5 wins out of 63 total games, which is about a 0.0806 proportion of wins or about a 8.06% win rate, traveled about 1,418 miles to play NCAA tournament games, and has only about $28,187 mean cost of attendance. For comparison, New York had 1 win out of 63 total games, which is about a 0.0161 proportion of wins or about an 1.61% win rate, traveled only about 731 miles, but has about $54,280 mean cost of attendance, which means that winning teams from Oklahoma traveled about 687 miles less than winning teams from New York, yet the Oklahoma teams won 4 more games than New York teams. In addition, students attending institutions in Oklahoma pay about $26,093, on average, less than students attending institutions in New York.

Now that we have seen, from quantitative data, some inconsistencies with the data that seems to refute the idea that there is a relationship between the average cost of attendance per state and either total travel distance per state or each state's proportion of wins, let's see if we can determine a relationship using the concept of linear regression and the least squares line, which is detailed in the theory portion. Here is the linear regression plot in blue and the residuals plot in orange:



As we can see from the above plot, the data does not look linear, so a linear regression model would not be appropriate. When performing the linear regression, we found that the r-squared value is 0.071554, indicating a non-ideal fit. Additionally, we do not see nearly normal residuals and inconstant variability, all of which do not meet the conditions for the least squares line.

Now, we wanted to see if there was any relationship between proportion of wins and mean cost of attendance. The following plots show their relationship:



As we can see from the above plot, the data does not look linear, so a linear regression model would not be appropriate. When performing the linear regression, we found that the r-squared value is 0.118866, indicating a non-ideal fit. Additionally, we do not see nearly normal residuals and inconstant variability, all of which do not meet the conditions for the least squares line. So, we can see that there is no linear relationship between the average cost of attendance per state and either total travel distance per state or each state's proportion of wins.

Theory

Summary Statistics

Summary statistics help us in seeing general trends that exist in our data. Example of summary statistics include mean, median, max, min, standard deviation etc. We used histograms, boxplots and Q-Q plots to help us graphically visualize these summary statistics.

The mean is given by the formula $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$. This sample statistic gives us one number that indicates what the variable would be for the entire population. The variance is another sample statistic that helps us look at the spread of the data, it is given by the formula $Var(X) = \frac{1}{n}\sum_{i=1}^{N}(X_i - \mu)^2$. The standard deviation is given by the square root of the variance.

For standard normal curves that are unimodal and symmetric, using chebyshev's theory, we know that 68% of the area under the curve corresponds mean ± 1 standard deviation. 95% of the area under the curve corresponds to the mean $\pm$ 2 standard deviations and 99.7% of the area under the curve corresponds to the mean $\pm$ 3 standard deviations.

The area under the curve can be given by the formula:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\Pi}} e^{\frac{-x^2}{2}}$$

For a Q-Q plot, the p-th quantile of the random variable X is defined as any number q satisfying $P(X \le q) \ge p$ and $P(X \ge q) \ge (1-p)$ with q sometimes denoted with $F^{-1}(p)$. The sample quantile are based on order statistics. For standard normal distribution, the q-th quantile is Z_q, where $\Phi(Z_q) = q,\ 0 < q < 1$. Examples of quantiles are the median, upper and lower, which correspond to 0.5, 0.25, 0.75 quantiles respectively. These are also the quantiles used for boxplots.

To find the sample quantiles of data $X_1 .....X_n$, we must first start by ordering the data from smallest to largest. X_k is then considered to be the $\frac{k}{n+1}$ th sample quantile.

You can compare two samples by comparing each corresponding ordered data point. For example we will expect $X_k \sim Y_k$. The Q-Q plot provides a graphical means of comparing the data distributions with the normal. If the plotted point lies on the line, then the data approximately has a normal distribution. If there is an upward curve in the plot, then the distribution has a long right tail in comparison to the normal, while a long left tail is indicated by a downward curve to the left. We can see the granularity of the data as stripes on the plot. And the bimodality of the data can be seen as a curved middle section of the plot.
Identical plots will have an intercept of 0, with a slope of 1. If the two distributions have the same shape, but different means, standard deviation, then the plot will be linear, with a different slope and intercept.

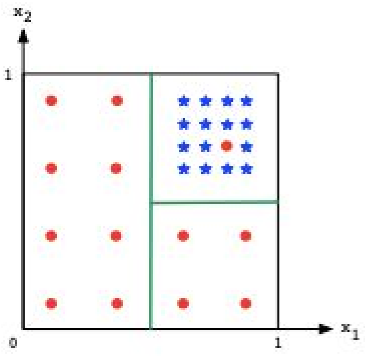
Decision Trees

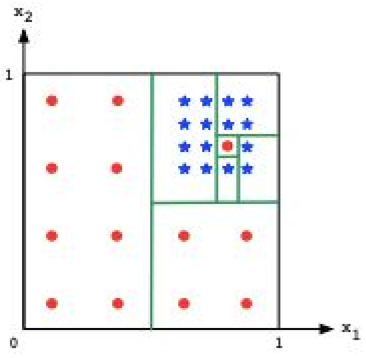
Decision trees are used mainly for classification and regression problems. Decision trees mimic human level thinking making it easy to understand data and make good interpretations for our predictions. Each node in the decision tree represents a feature, and each branch represents a decision, while each leaf node represents the final category predicted.

The greedy algorithm to build a tree top down is as follows:

```
Start with a single node containing all data points
Repeat:
    Look at all current leaves and all possible splits
    Choose the split that most decreases the uncertainty in prediction
```

We will consider the label + to be p fraction of the points and - label to be (1-p) fraction of the points.
The uncertainty in prediction is given by the misclassification rate (min(p,1-p)), and the gini index (2p(1-p))

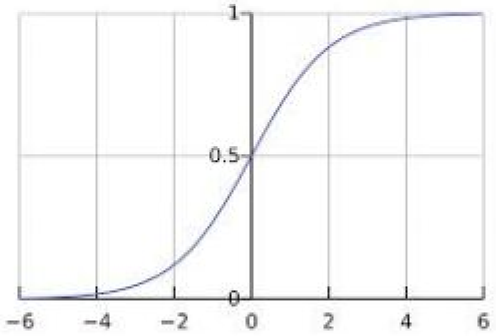




We must ensure that we are not overfitting our data. In the above picture we can see that on the right we are overfitting the model to noise seen in our data set. This will reduce the accuracy of our model on unseen data, since that one point may be an outlier in our data, yet we are learning the point in our model.
The main benefits of a decision tree is that it can accommodate any type of data (real, boolean, categorical). It can also accommodate any number of classes and fit any data.

Logistic Regression
Logistic regression works similar to linear regression, but instead of outputting a real number, we output the class it belongs to. We do this by using a squashing function for the real number output from the linear regression. We use a sigmoid function to do the same.

$$s(z) = \frac{1}{1+e^{-z}}$$



$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w*x + b)}}$$

Hypothesis testing
A statistical hypothesis is an assumption about a population parameter. We start of by creating a null hypothesis which is a belief of our population parameter, and an alternative hypothesis which the opposite of our null hypothesis. We then simulate to calculate the observed count of the population parameter, if the observed proportion is below a certain significance level (usually 0.05), then we will reject our null hypothesis for our alternative hypothesis, else we will continue to believe in our null hypothesis. Hypothesis testing can result in two main types of decision errors, notably type 1 and type 2 errors. Type 1 errors is when the null hypothesis is true, yet we reject it, similar to a FP, while a type 2 error is when we fail to reject the null hypothesis that is false, similar to false negative.

Chi-Squared Tests

A chi-squared test is a form of hypothesis test that compares a sampling distribution with a null distribution using the formula:

$$\mathrm{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$$

where ‘O’ refers to observed counts for each category and ‘E’ refers to expected counts for each category. The chi-squared statistic has an associated degrees of freedom determined by (# rows - 1)(# columns - 1). By these two metrics we can find a corresponding probability that that sampling distribution or one more extreme would occur from the expected model.

Haversine Distance Formula

The haversine distance formula calculates the shortest distance between two points on a sphere, which in this case is Earth, using their latitude and longitude coordinates measured along the surface. Here is the following equation for haversine distance:

$$\text{haversin }\alpha = \text{haversin }(\theta_1-\theta_2) + \cos\theta_1\cos\theta_2\,\text{haversin }(\phi_1-\phi_2)$$

Conditions for Least Squares Line

1. Linearity
2. Nearly Normal Residuals
3. Constant Variability

The definition of linearity is that the relationship between the explanatory and the response variable should be linear. To confirm whether linearity exists, we need to check using a scatterplot of the data or a residuals plot.

The definition of nearly normal residuals is that the residuals, like in a residual plot, should be nearly normal. This condition may not be satisfied when there are unusual observations, such as outliers, that do not follow the trend of the rest of the data. To confirm nearly normal residuals, check using a histogram or normal probability plot (Q-Q plot) of residuals.

The definition of constant variability, or homoscedasticity, is that the variability of points around the least squares line should be roughly constant. This implies that the variability of residuals around the 0 line should be roughly constant as well. To confirm constant variability, check using a histogram or normal probability plot of residuals.

Ordinary Least Squares

The ordinary least squares (OLS) approach to regression allows us to estimate the parameters of a linear model. The goal of this method is to determine the linear model that minimizes the sum of the squared errors between the observations in a dataset and those predicted by the model.

Permutation Tests

A permutation test, also known as a randomization test or an exact test, is a type of statistical significance test, in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. For a permutation test, we define N = n + m, and let v = ($v_1, v_2, \ldots , v_N$) be the ordered and grouped vector of the observed values a, b. In addition, let g = $g_1, g_2, \ldots , g_N$) be the vector that indicates the observations' original labels. There are $(N\ c\ n) = \frac{N!}{n!m!}$ possible g vectors. Under the permutation lemma, under $H_0 : F = G$, the vector g has probability $\frac{1}{(N\ c\ n)}$ of equaling any one of its possible values. Let g^* indicate any one of the $(N\ c\ n)$ possible vectors of g, and $\theta^* = \theta^*(\ g^*\ )$. The achieved significance level (ASL) of the permutation is defined as the following: $ASL_{perm} = Prob_{perm}\{\theta^* >= \theta\ \} = \#\{\theta^* >= \theta\ \}\ /\ (N\ c\ n)$

Advanced Analysis: The Model

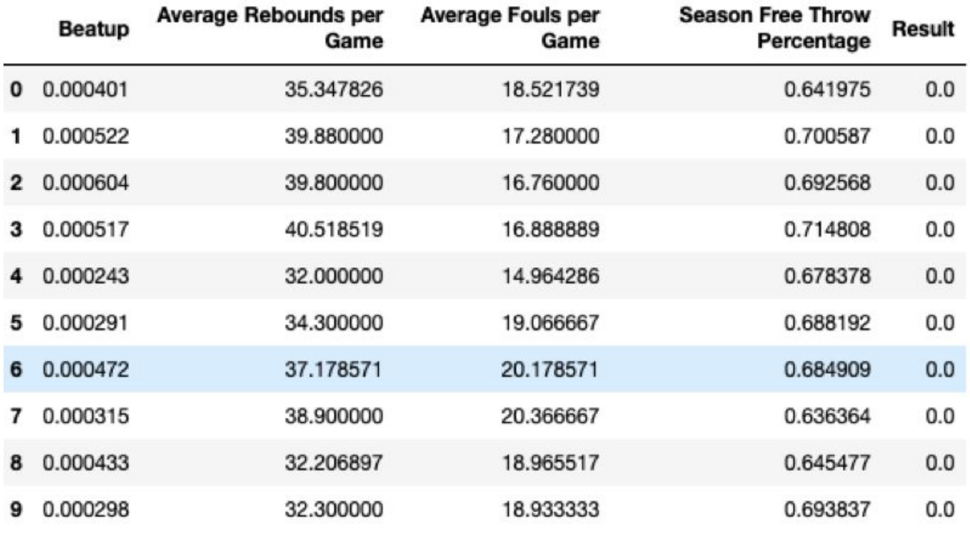
Our advanced analysis is focused on building our predictive model. In creating our model, we sought to accurately classify teams into one of the four categories:

- Didn’t Make NCAA tournament (corresponds to 0 in Result column of the table below)
- Make NCAA tournament but didn’t make the Final 16 (corresponds to 1 in Result column)
- Made the Final 16 but didn’t make the Final 4 (corresponds to 2 in Result column)
- Made the Final 4 (corresponds to 3 in Result column)

Based on our findings in the basic analysis, our group decided to include the following features in our final model:

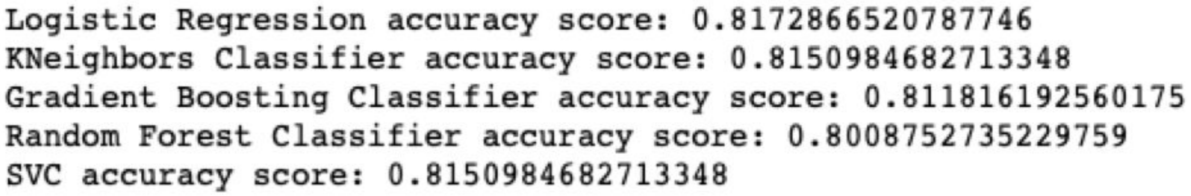
- Beat-up Metric (produced in “Advanced Metrics” section)
- Average Rebounds per Game
- Average Fouls per Game
- Season Free Throw Percentage

The first 10 entries of the table shown below contains the metrics described above for each team in each NCAA season since 2010:



	Beatup	Average Rebounds per Game	Average Fouls per Game	Season Free Throw Percentage	Result
0	0.000401	35.347826	18.521739	0.641975	0.0
1	0.000522	39.880000	17.280000	0.700587	0.0
2	0.000604	39.800000	16.760000	0.692568	0.0
3	0.000517	40.518519	16.888889	0.714808	0.0
4	0.000243	32.000000	14.964286	0.678378	0.0
5	0.000291	34.300000	19.066667	0.688192	0.0
6	0.000472	37.178571	20.178571	0.684909	0.0
7	0.000315	38.900000	20.366667	0.636364	0.0
8	0.000433	32.206897	18.965517	0.645477	0.0
9	0.000298	32.300000	18.933333	0.693837	0.0

To determine the best regressors/classifiers to use, we selected five common ones, namely Logistic Regression, K- Nearest Neighbors, Gradient Boosting Classifier, Random Forest Classifier and SVC. The accuracy scores for each of the regressors/classifiers are shown below:



```
Logistic Regression accuracy score: 0.8172866520787746
KNeighbors Classifier accuracy score: 0.8150984682713348
Gradient Boosting Classifier accuracy score: 0.811816192560175
Random Forest Classifier accuracy score: 0.8008752735229759
SVC accuracy score: 0.8150984682713348
```

As seen above, each of the classifiers gives an accuracy score of around 0.80, 0.81, with the logistic regression classifier having the best accuracy score of 0.8173. As a constant model predicting all 0’s would have an accuracy score of 0.815, our model only minutely improves on the constant model.

Conclusion

From Scenario 1, we looked at the performance of both home and away teams using data from the 2010-2017 seasons. We found that, for the regular season, the distribution of point differences between teams to be right skewed, which makes sense since most games would end up as close match ups. Specifically, we found that this histogram is unimodal, with a mode of around 5. Additionally, the distribution has a long right tail, making it right skewed. For the knockout rounds, we saw a similar distribution in that we got a unimodal histogram, with a mode of around 5. The distribution has a long right tail, making it right skewed. Next, we found that, for total points scored during the regular season and NCAA tournament, total points had normal distributions. Then, we used choropleth maps to determine that states like New Hampshire and Rhode Island have closer games, while games in Oklahoma and West Virginia are more one sided. Then, we found that, from our boxplots, there is no indication of whether attack or defense is more important during the regular season. For the NCAA tournament, however, we can see that home losing sides have significantly more points scored than away losing sides.

For Scenario 2, we performed feature engineering on a team's seed to help develop a model to predict tournament success. To determine tournament performance, we looked at the round at which a team was defeated. Then, we found that the round of 64 is the most frequent tournament placing out of all 16 seeds. Next, we ran chi-square tests, which would express how different each observed counts were from average expected placements. From our analysis, no observed distribution of counts of each round the team is out is modeled after the uniform distribution of expected counts.

From Scenario 3, we looked at the differences in the "hard-work" metrics, such as fouls and rebounds, for teams admitted to the NCAA Tournament. Specifically, to detect a statistical difference amongst attributes, we conducted various t-tests in order to decide whether or not these statistics are the difference maker when it comes to whether or not a team makes the Final Four. One interesting insight while performing data analysis is that the University of Connecticut (UConn) has made every Final Four since 2008, making UConn a bit of a dynasty institution. Using bar plots, we determined that teams that make it further in the NCAA tournament tend to pick up a lower amount of fouls than those teams that exit earlier in the tournament. By using hypothesis testing, we find that there exists a statistically significant difference in the average number of fouls committed amongst the final four teams and the rest of the field. Again, using hypothesis testing, we find a statistically significant difference in the average number of steals per game amongst the final four teams and the rest of the field does not exist. Furthermore, we find that teams that are able to rebound at a higher rate, on average, are able to move deeper into the tournament and have a better chance of reaching the Final Four.

From Scenario 4, we wanted to determine if the distance traveled to play an NCAA tournament game had any effect on a team's performance in the game, particularly if distance traveled has any effect on a team's proportion of wins. Once we found the haversine distances of each winning and losing team to their respective game location, we plotted them and found that a majority of games and institutions are from the Midwest to East Coast regions. Initially, we found inconsistencies in a potential relationship between travel distance and proportion of wins. Next, we found that there is no linear relationship between these features. However, an interesting insight is that about 65% of winning teams in both the 2010 and 2017 seasons traveled less than the losing teams, suggesting that a shorter travel distance would favor a team to win over a longer travel distance.

From Scenario 5, we sought to create a feature that quantifies how “beat-up” a team is at the end of their season. For this scenario, we used metrics, such as overtime, days of rest, median fouls against, and median travel distance. To determine if the above variables are useful in predicting a team’s success, we used a Random Forest Classifier to find the proportion each of the above features contributed to predicting team success. One insight was that while there is much overlap in the beat-up scores among the different values of the Result variable, it appears that teams that didn’t make the NCAA tournament are more likely to have higher beat-up scores. Then, we conducted a permutation test to determine if there was a statistical significance in the beat-up means, and we found that there is such a significance.

From our additional analysis, we wanted to see if there is any relationship between the average cost of attendance per state and either total travel distance per state or each state’s proportion of wins. Similarly to travel distance, we found that there is no linear relationship between mean cost and proportion of wins. From there, we created a model to help us predict the outcome of the March Madness bracket using the features that we engineered throughout our scenarios. We built a model that could output an accuracy of about 80%.

Works Cited

https://simplemaps.com/data/us-cities
https://www.latlong.net/place/hesburgh-library-notre-dame-indiana-22045.html
https://www.latlong.net/place/beaver-stadium-university-park-pa-usa-23019.html
https://nces.ed.gov/ipeds/use-the-data
https://www.math.ksu.edu/~dbski/writings/haversine.pdf
https://pdfs.semanticscholar.org/f158/7608c2c6cb8fd26b50f20943849344dee892.pdf
https://seeing-theory.brown.edu/regression-analysis/index.html
http://www.ncaa.org/championships/statistics/ncaa-womens-basketball-championship-tournament-records
https://www.kaggle.com/c/womens-machine-learning-competition-2019
https://www.tau.ac.il/~saharon/StatisticsSeminar_files/Permutation%20Tests_final.pdf