# Final Project Step 3

Ramsey King

2021-06-05

## Introduction

The problem that I have is that I am overly competitive. Every year, my family competes together in a bracket for March Madness, with the winner getting to be treated by the family to a lunch of their choosing. For many years, I have been the far and away winner of this pool (competition). This year, however, I was unsuccessful in winning the family pool. I would like to have my title back. In wanting to return to glory, I have decided that I will use exploratory data analysis, statistics, and R programming to see if patterns exists that will help me choose the Final Four teams of the men's basketball tournament. By consistently and correctly predicting the Final Four teams of the tournament, it will all but ensure that I continue my reign as the family basketball bracket picking champion.

Why would anyone besides me think that this is important? The ability to pick teams for the college basketball tournament has turned into a lucrative exercise. Even ESPN will award a winner $1 million if they are able to choose correctly every game during the tournament. By the way, there is a 1 in 9,223,372,036,854,775,808 chance of doing this, but it's greater than 0. By being able to choose the teams consistently that will reach the Final Four, there is real money to be had. If you do a Google search on "March Madness Bracket Contests," you will see grand prizes of cash money and trips to Hawaii and Las Vegas.

This is a data science problem because it includes a lot of probability, and there are many metrics that are kept with college basketball teams throughout the course of a season upon which statistical analysis can be performed.

## Research questions

1. Is there a pattern that exists that will predict the Final Four Teams?
2. What are the most important metrics that will help determine the success of a college basketball team in the tournament?
3. What are some metrics that may not seem important initially but then prove to have statistical significance?
4. If a pattern exists, will it truly be causation, or is it just a random happening?
5. What is the smallest number of predictors that can produce the best model?

## Approach

I will gather data from 1985-2016 or 2017 to see if a predictive pattern exists to select the Final Four teams.

## How your approach addresses (fully or partially) the problem.

My approach will address the problem because I will have historical data from the last 20 years on basketball teams that have made the tournament and how they have fared. Based on the teams that have made the Final Four, there may be patterns that exists (number of wins, strength of schedule, ranking, etc.) that well help predict how future teams will perform in the tournament.

## Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

Data sets that will be used are:

- Big_Dance_CSV
- ncaa-team-data
- NCAAInstitutions
- RPIStats

**Big_Dance_CSV data set:**

The Big Dance CSV data set was obtained from **Big_Dance_CSV.**

The original purpose of the Big Dance CSV data set was to provide "every NCAA tournament game result since 1985 (when the tournament was expanded to the 64 team bracket). The dataset contains the year, round (1-6), seed of the teams (1-16), region (1-4) and the scores."

**ncaa-team-data**

The ncaa team data data set was obtained from **ncaa-team-data**

The description of the data set is "I scraped data from sports-reference.com and made it tidy."

**NCAA Institutions**

The NCAA Institutions data set was obtained from Wikipedia. The data was collected and then put into a CSV file to be used for joining the other data sets together. The link to the Wikipedia file is **List of NCAA Division I institutions**

**RPIStats**

The RPIStats dataset was obtained from multiple sources and combined into one. For the years 1994 through 2017, the data was obtained from **College RPI**. For the years 1985 through 1993, the data was obtained from **RPI Archive**.

Because Big Dance CSV includes data going back to 1985 (which is the first season that the tournament went to 64 teams), I wanted to have RPI information going back that far as well, so both data sets will need to be joined together into one dataset.

## Required Packages

The packages that will be needed are ggplot2, dplyr, QuantPsyc, readExcel, and psych. I am sure that there will be other packages needed, but I am not aware that I will need them at this time.

## Plots and Table Needs

I will need to use histograms, scatterplots, and basic tables for to describe tabulated results.

## Questions for future steps

I need to learn how to join tables together in R. I am not quite comfortable with that at the moment. I will also need to figure out how to create other plot types outside of histogram and scatter plots to help visualize the data and findings. As I get more involved in the project, I am sure that I will discover other things that I do not know yet that I will need to figure out.

# Step 2

## How to import and clean my data

There has been (what I consider) a LOT of cleanup and manipulating of the data to work my way to the final data set. First of all, I have changed all the csv files (Big_Dance_CSV, ncaa-team-data, NCAAInstitutions,

and RPIStats) to xlsx (Excel) files. The reason I have done this is because there is functionality in xlsx files that are easier to work with than csv files, and xlsx files "remember" formulas better (at least that is what I have experienced so far). A lot of this cleanup and manipulation has been done within Excel, but there will be some manipulation that will take place in R. I will describe the data prep for the specific csv/xlsx files.

**Big_Dance_CSV data set:** For the Big_Dancd_CSV data set, the following changes were made:

- changed the file extension from CSV to XLSX (as mentioned above)
- added the following columns:
  - TeamW,
  - TeamWSeed,
  - TeamWCommonName,
  - TeamWID,
  - TeamL,
  - TeamLSeed,
  - TeamLCommonName,
  - TeamWLID

The TeamW/TeamL columns were added to help identify which teams won the game played on the row and which team lost.

The TeamCommonName columns were added to be able to connect to the NCAAInstitutions data set.

The NCAAInstitutions data set will be the data set to identify Teams between all the other data sets.

**ncaa-team-data**

- added common name and TeamID columns to be able to tie to the NCAAInstitutions dataset.

**RPIStats and RPI_Dataset** The data from rpi95_04 and RPIStats were combined together into RPI_Dataset. The RPI_Dataset contains RPI (Ratings Percentage Index) information for all teams from 1985 through 2016. The Ratings Percentage Index is a metric used in college sports to help determine a team's relative team strength. It is a metric that is used in ranking and seeding the basektball teams for the single elimination tournament.

## What does the final data set look like?

The final data set will be a collection of all the basketball teams that played in the NCAA single elimination championship tournament from 1985 to 2016 and will contain the following columns:

- CommonNameSchool (school name that will link to NCAAInstitutions dataset)
- W (Wins)
- L (Losses)
- WL (Winning %)
- srs (Sports Reference Statistic - give definition of this)
- sos (Strength of schedule)
- ncaa_result (How the team fared in the NCAA single elimination tournament)
- ncaa_numeric (a numerical representation of how the team did in the NCAA tournament)
- TeamID (A team ID nummber to link to the NCAAInstitutions dataset)
- Year
- Seed
- TeamW, TeamL (specific game info in big_dance_csv data set)
- TeamWSeed, TeamLSeed
- RPI
- OWP, OOWP (from 2003 through 2016)

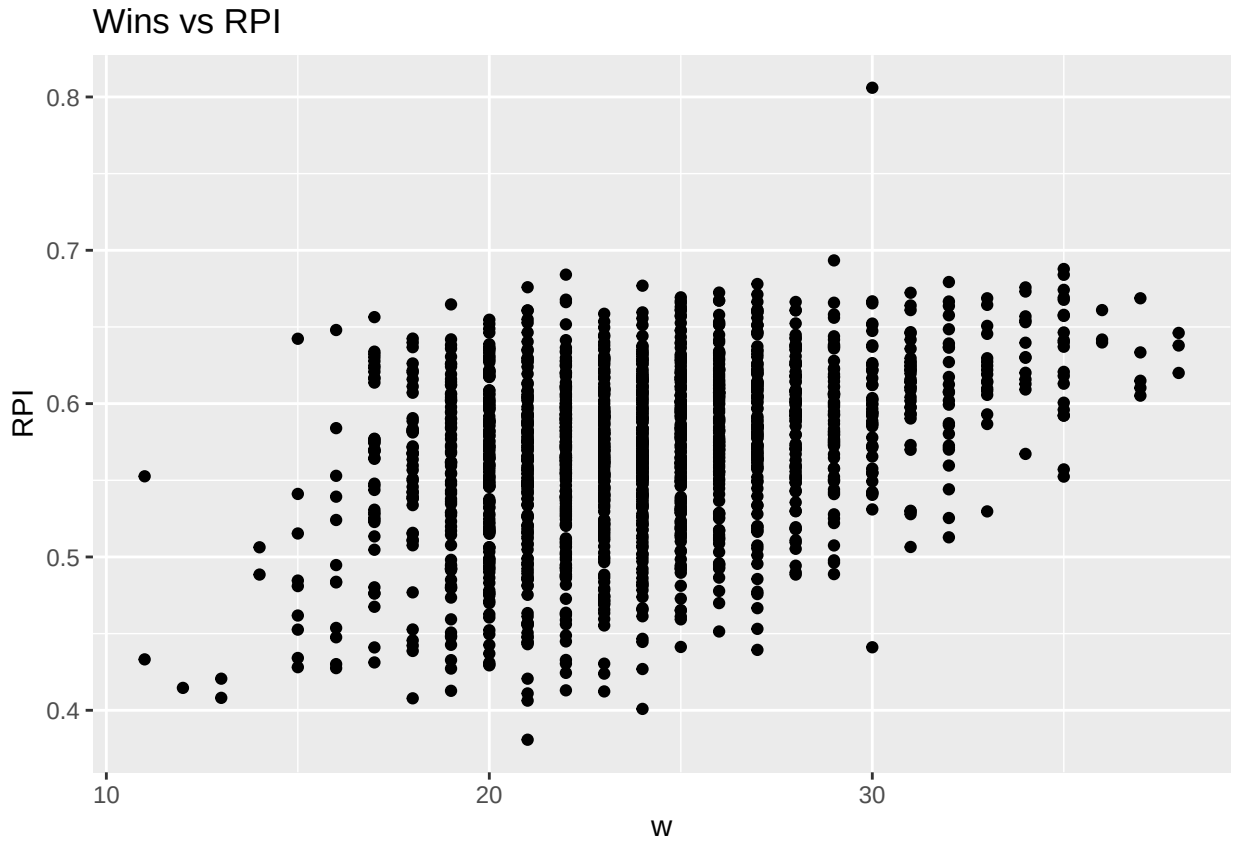
## Questions for future steps.

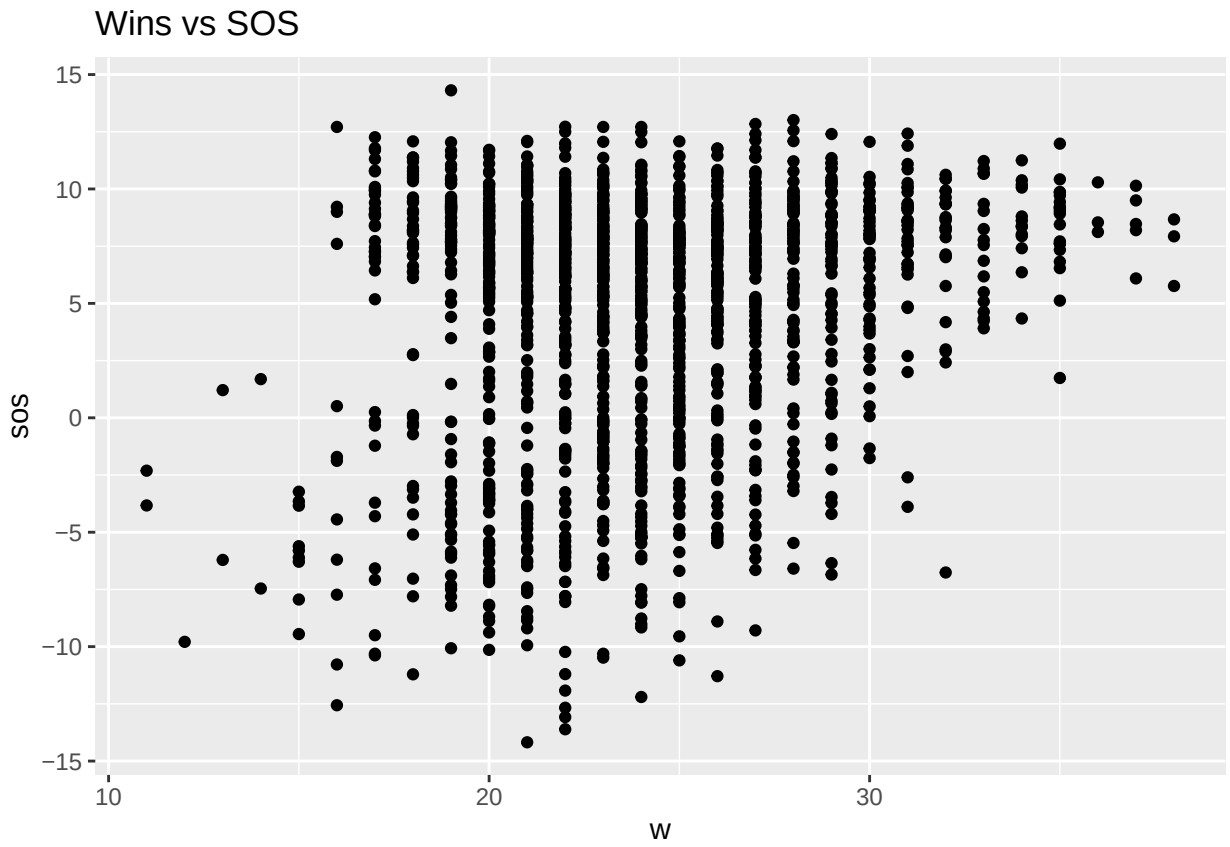
### What information is not self-evident?

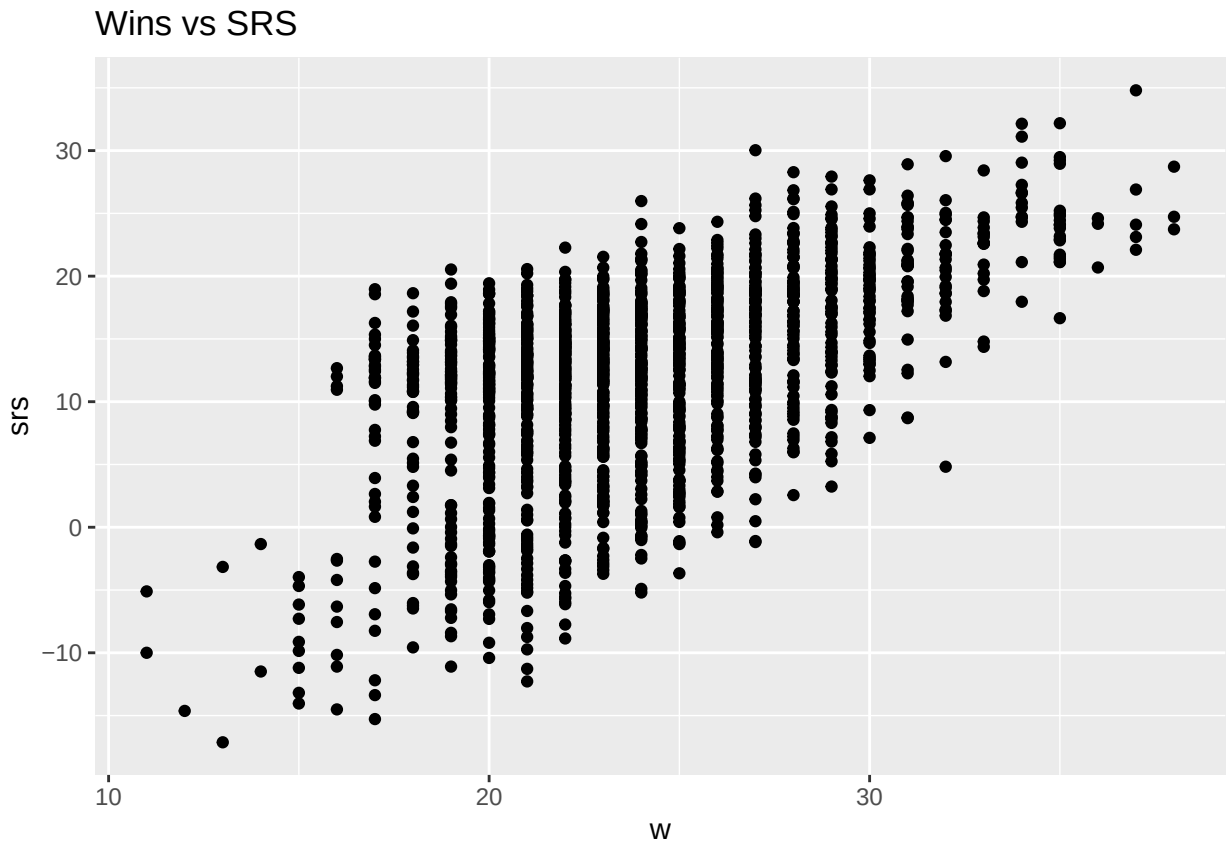
The information that is not self evident is if there is a pattern that exists that will predict the final four teams. I also may need to think about the seeding to make sure that I only choose teams that can make the final four (don't want two teams from the same region making the final four, that's impossible)

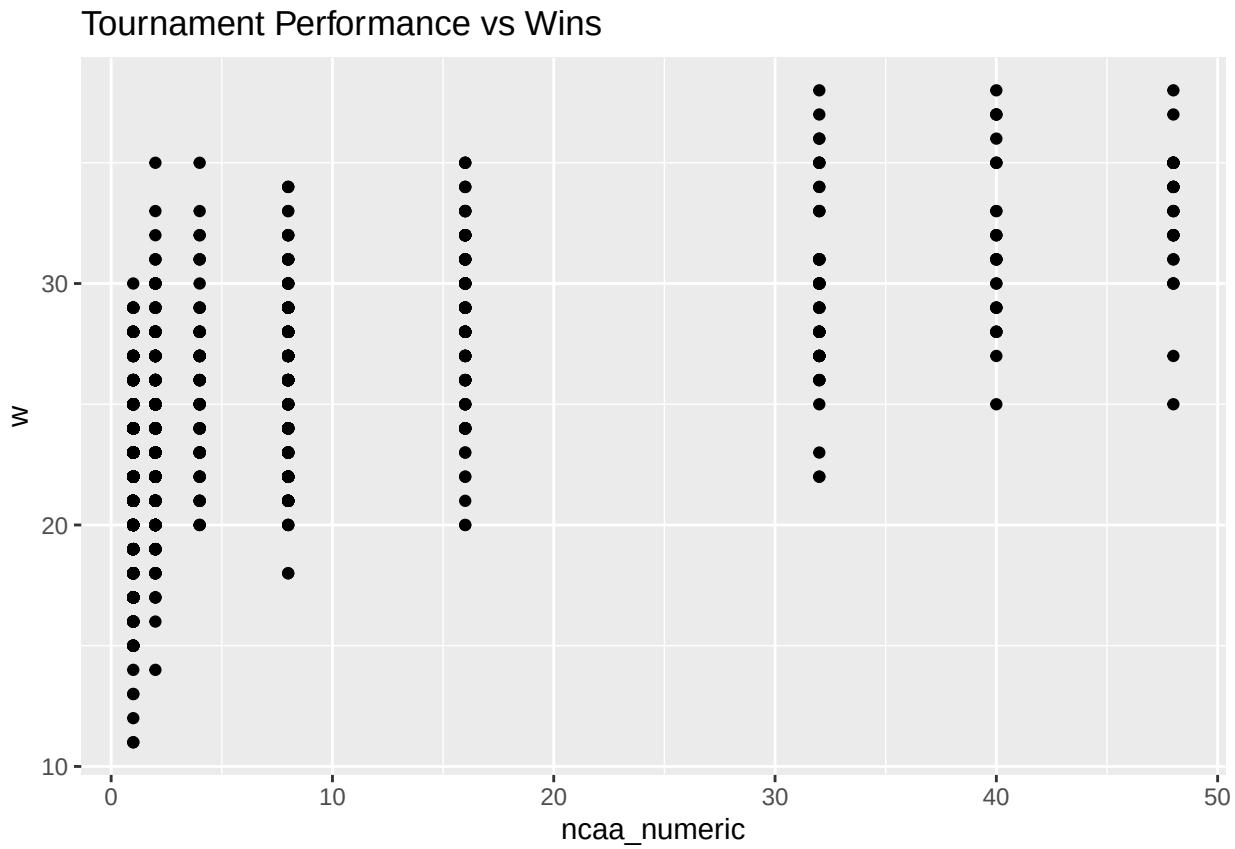
### What are different ways you could look at this data?

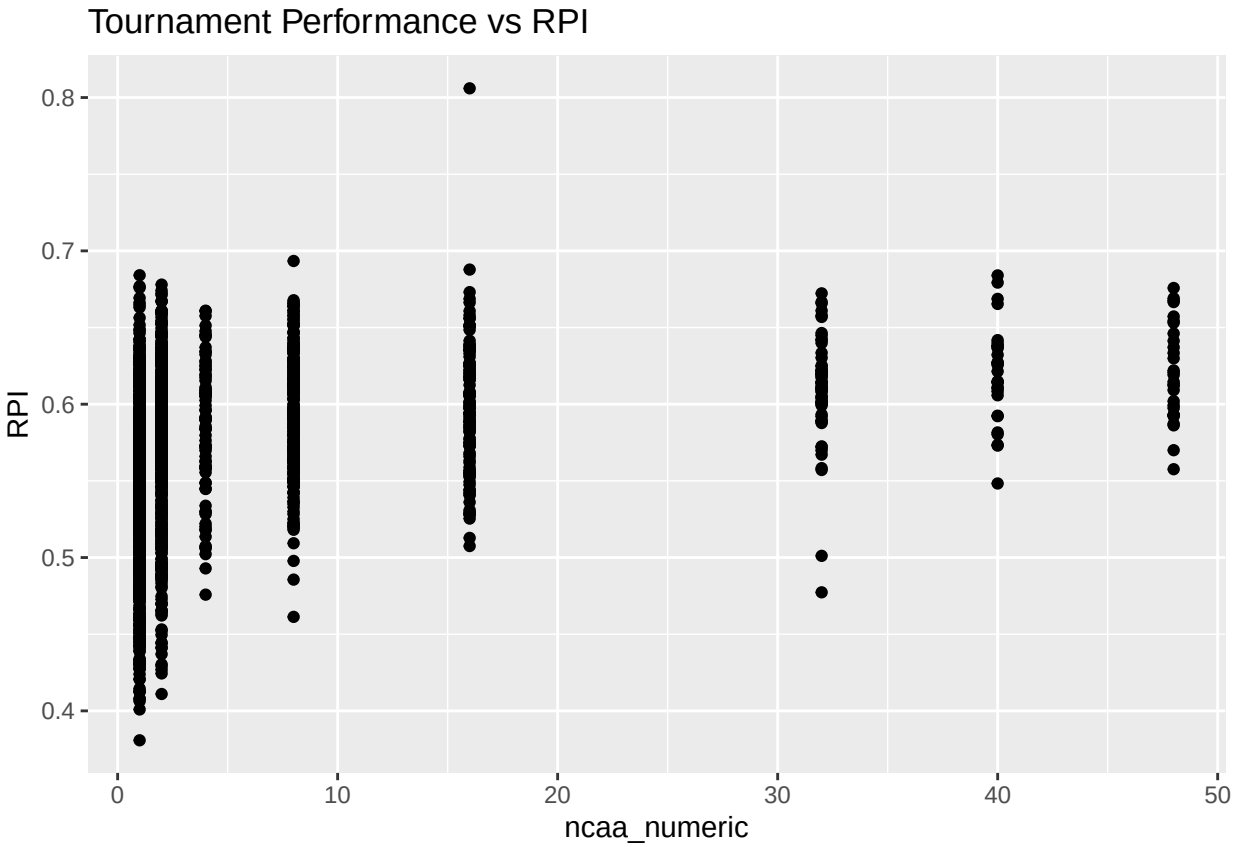
I plan on doing scatterplots for all metrics that I am interested in against the teams that made the final four. For example, Wins vs. RPI, Wins vs. SOS, Wins vs. SRS, etc.

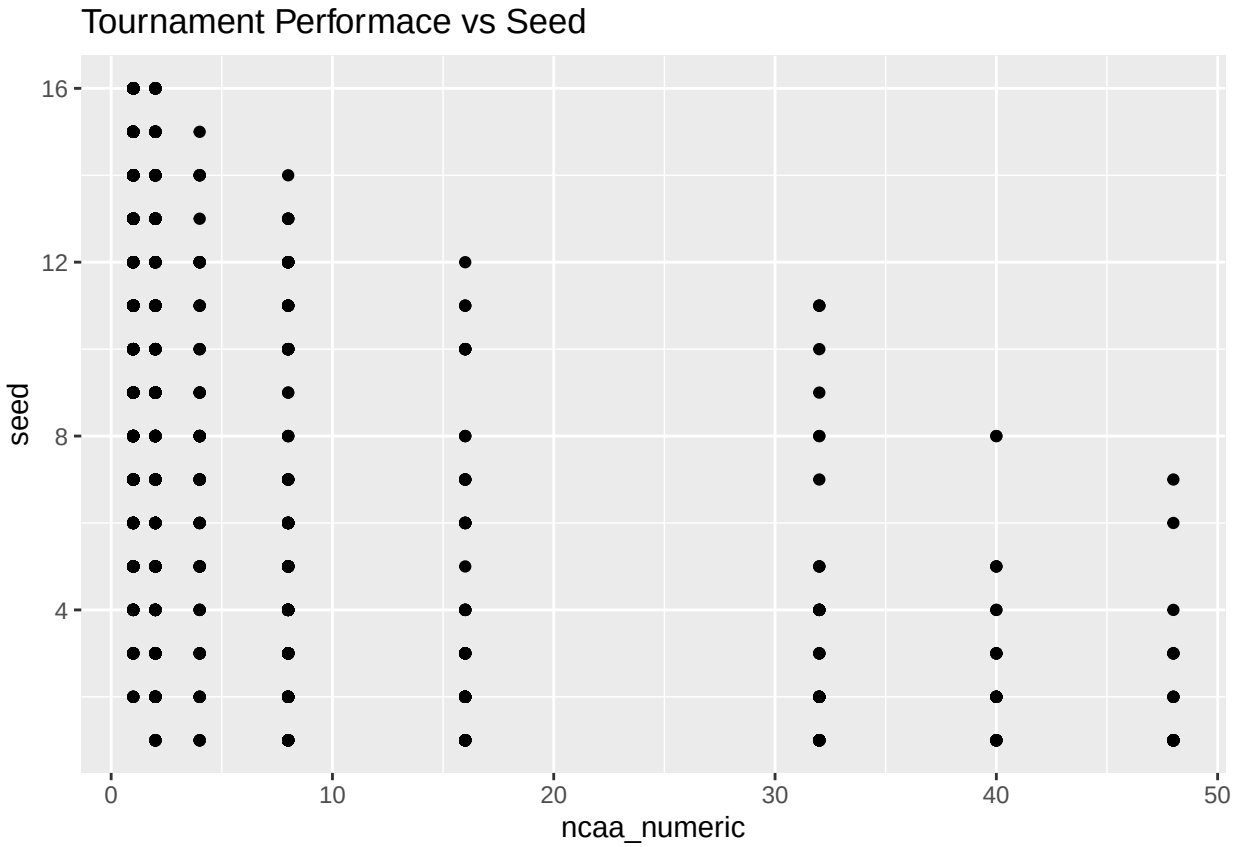



Wins vs SOS
15
10
5
0
−5
−10
−15
sos
10
20
30
w


Wins vs SRS
srs
30
20
10
0
−10
10
20
30
w

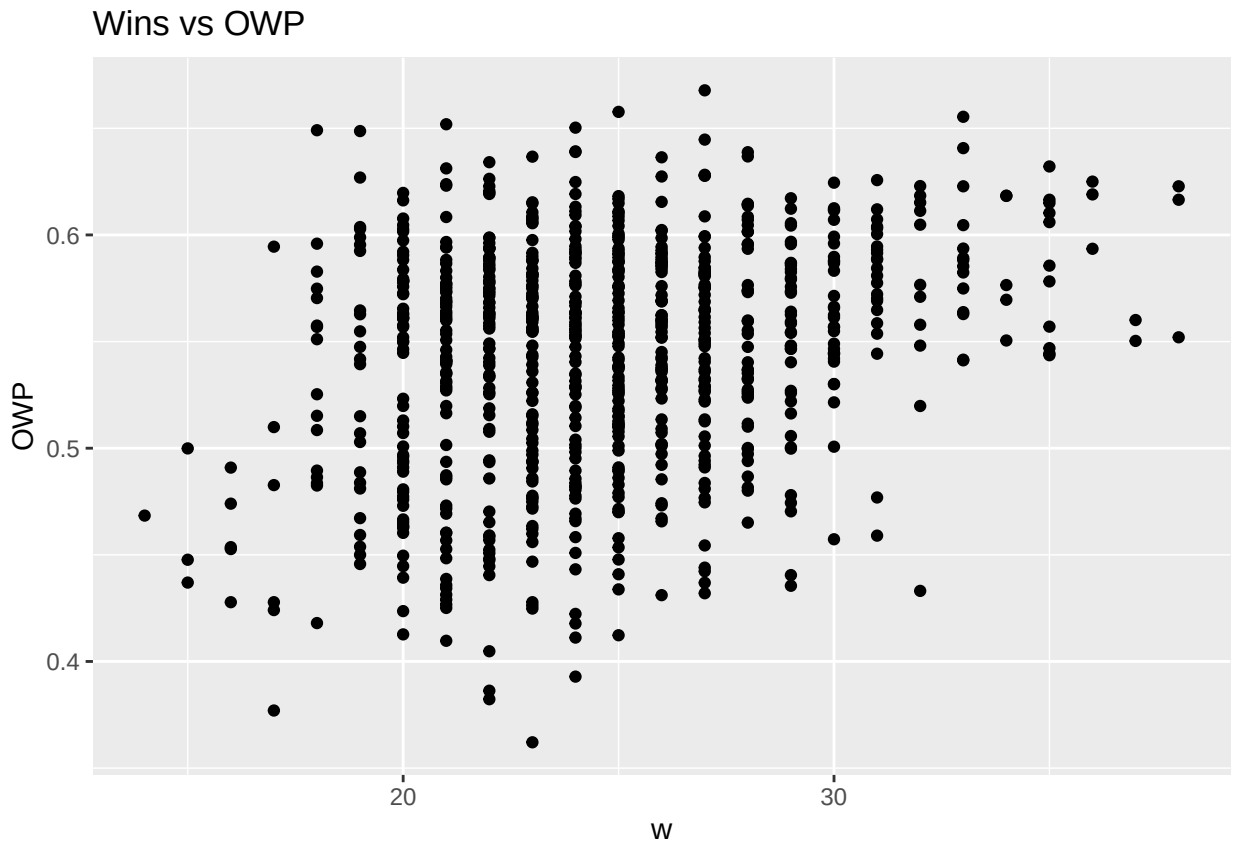

Tournament Performance vs Wins
w
30
20
10
0
10
20
30
40
50
ncaa_numeric

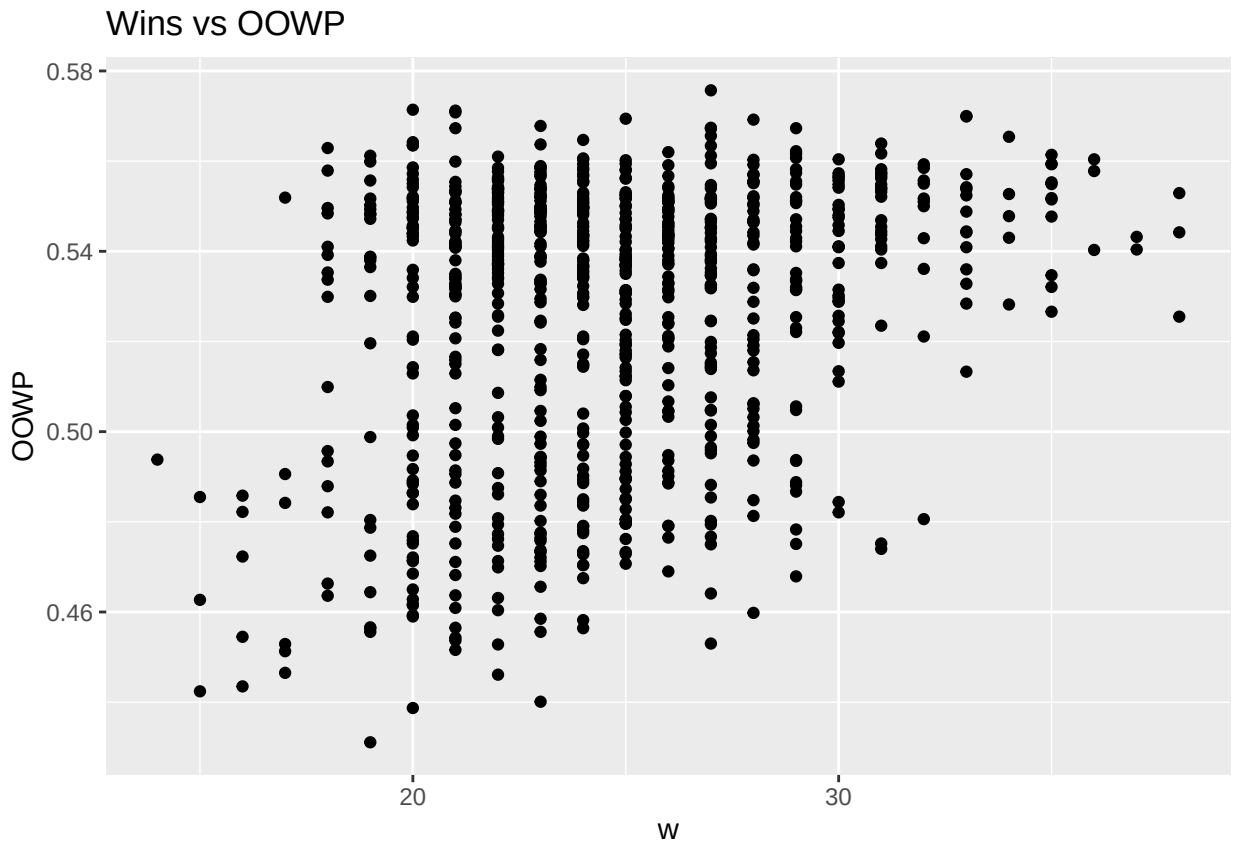

Tournament Performance vs RPI
RPI
0.8
0.7
0.6
0.5
0.4
0
10
20
30
40
50
ncaa_numeric

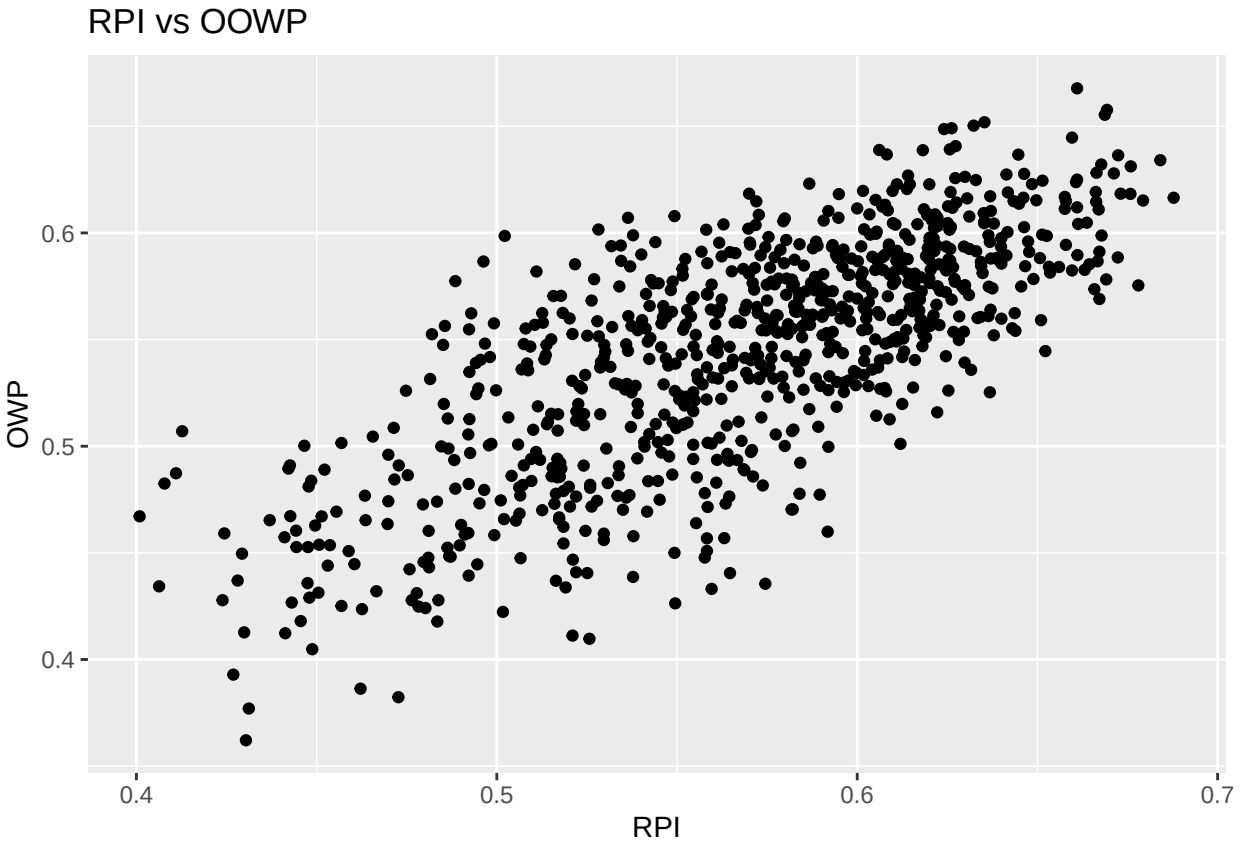


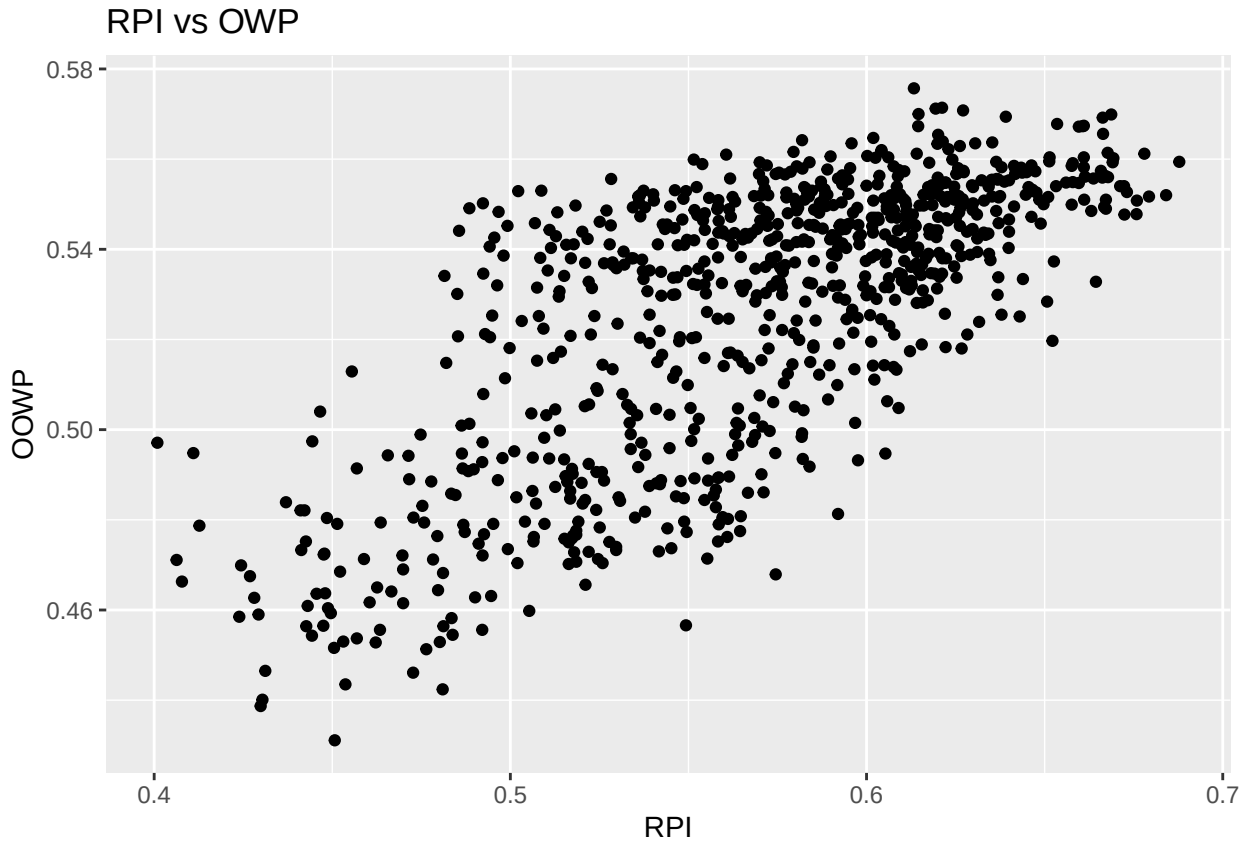

### How do you plan to slice and dice the data?

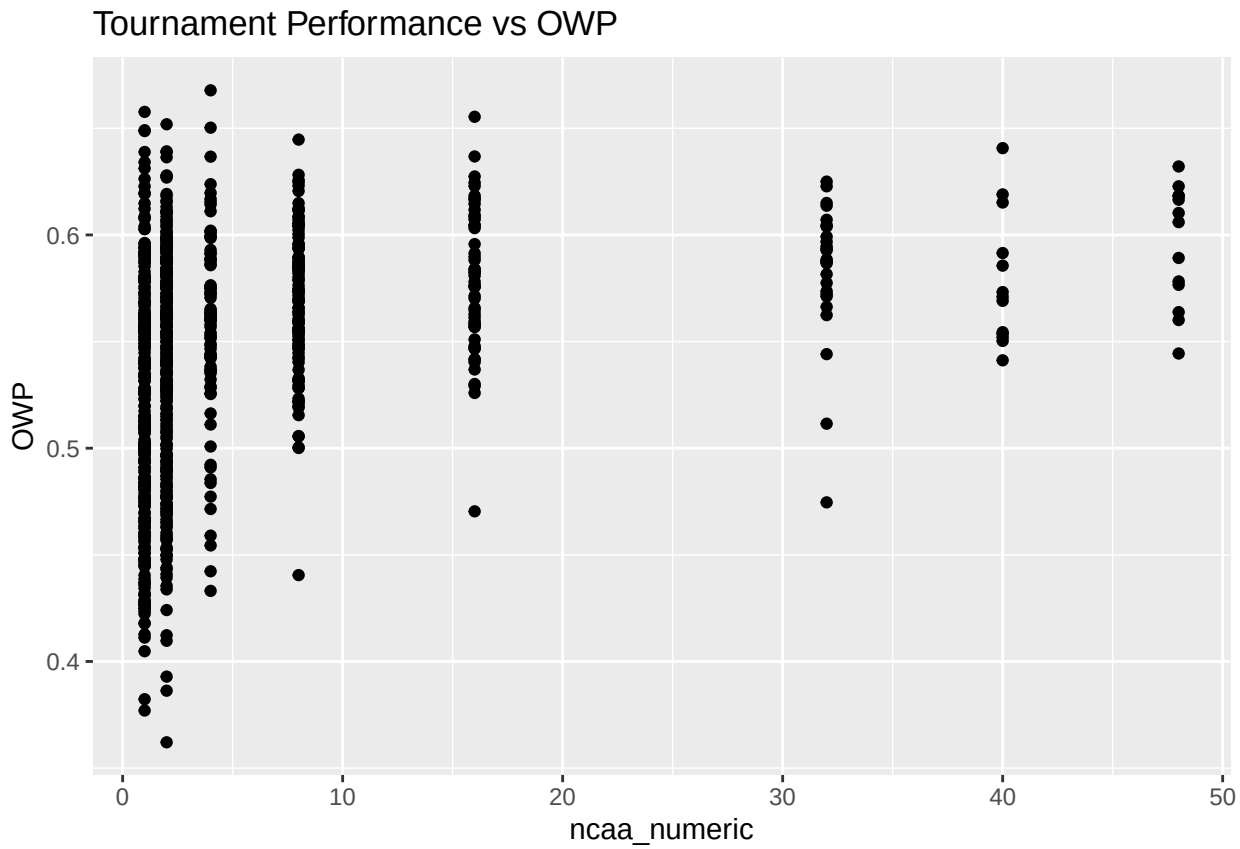
One way that I can slice the data is that I can create a subset of the data set from 2003 forward. This will allow me to include the OWP & OOWP metrics in the predictions.


Wins vs OWP
OWP
0.6
0.5
0.4
20
30
w


Wins vs OOWP
0.58
0.54
0.50
0.46
OOWP
20
30
W


RPI vs OOWP
OWP
0.6
0.5
0.4
0.4
0.5
0.6
0.7
RPI


RPI vs OWP
0.58
0.54
0.50
0.46
OOWP
0.4
0.5
0.6
0.7
RPI

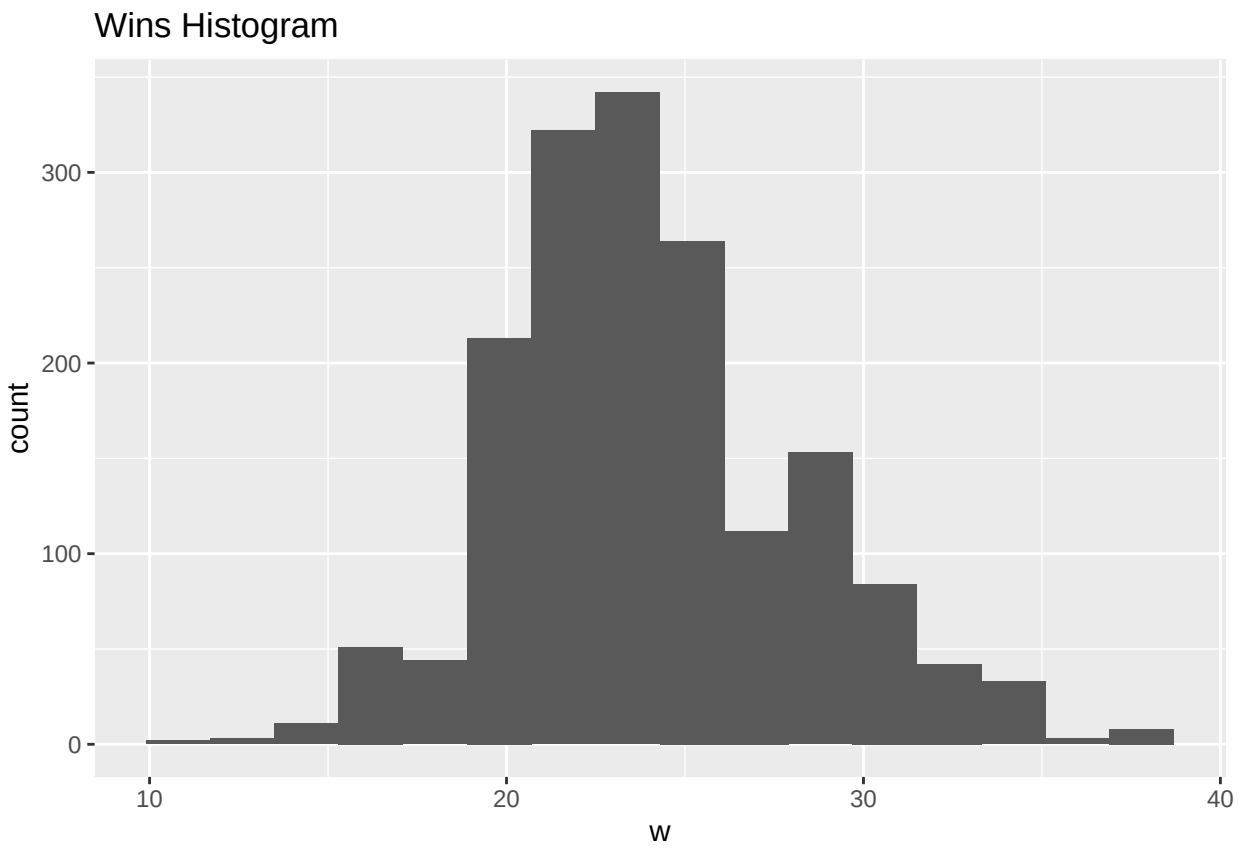


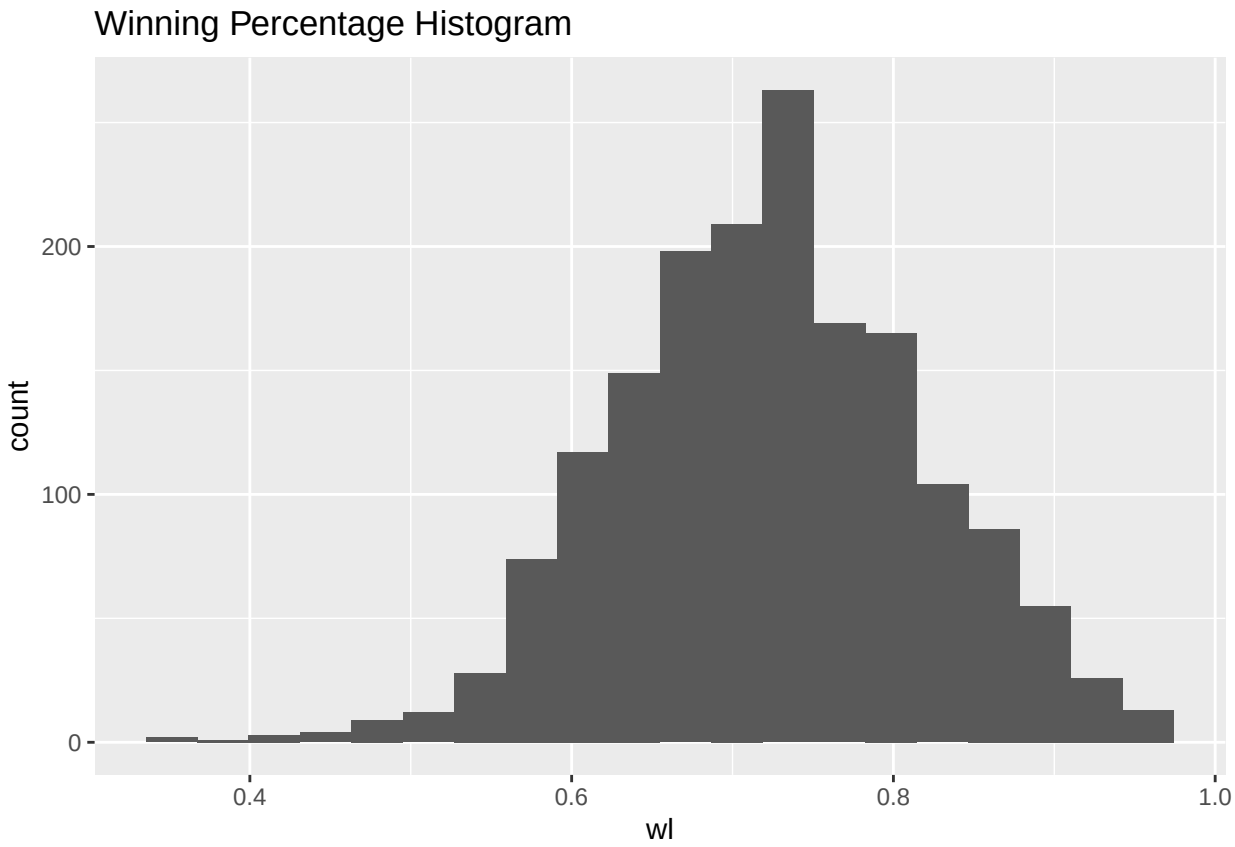

### How could you summarize your data to answer key questions?

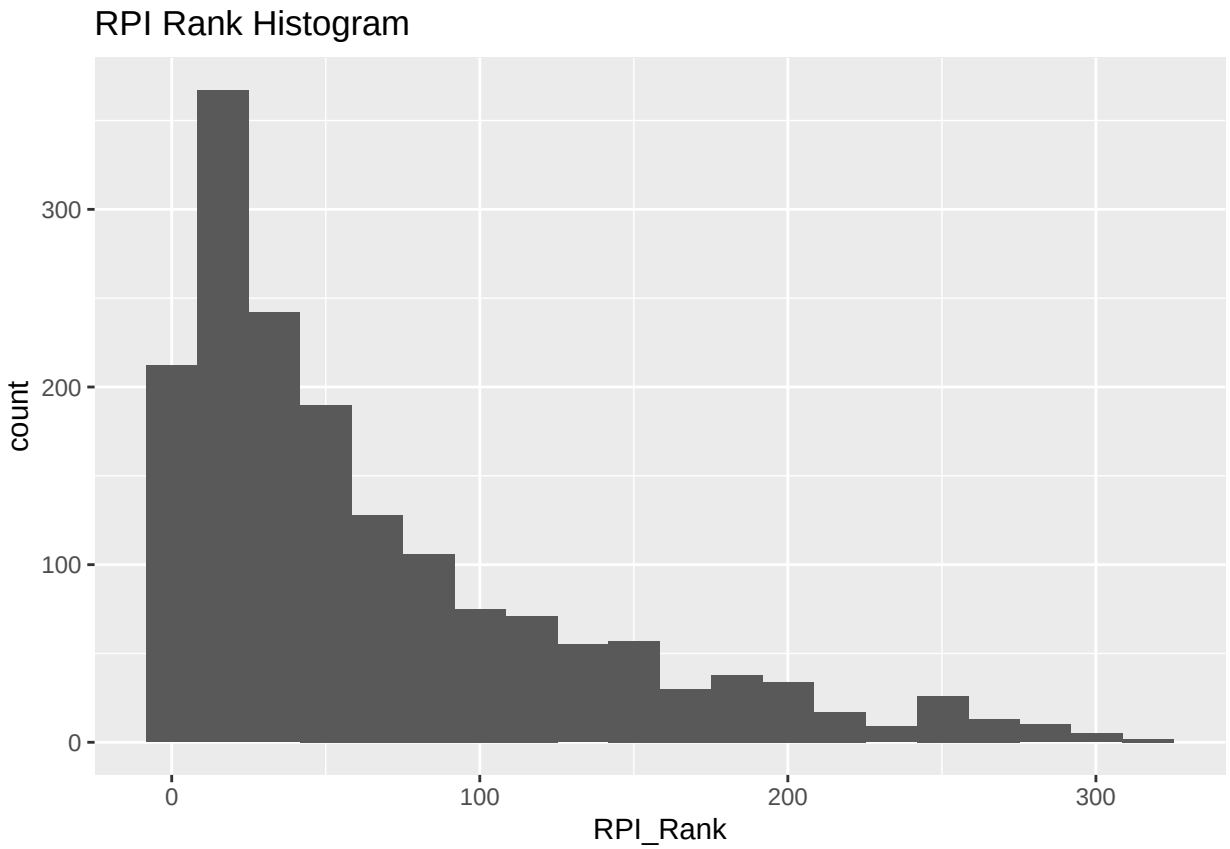
Whatever pattern exists, if any, that will be used to answer the five questions will be presented in this section.

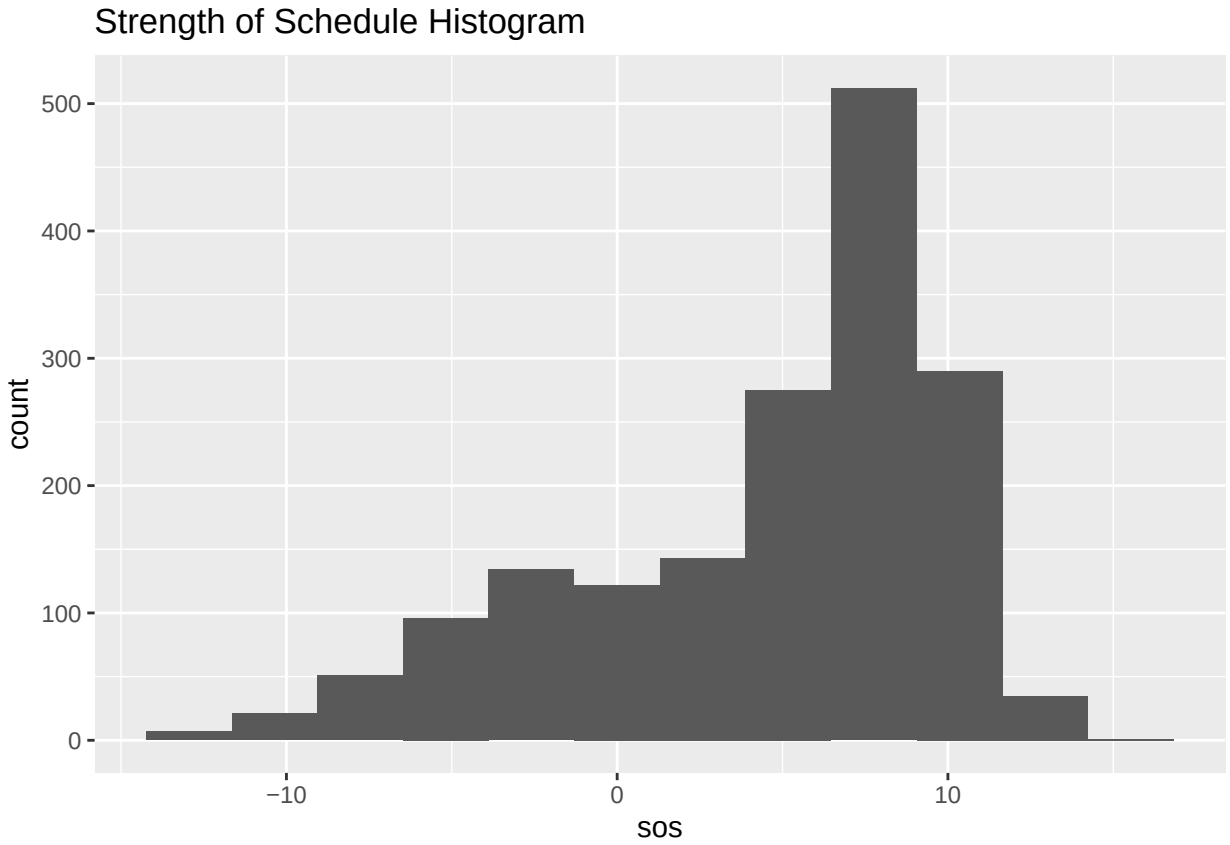
### What types of plots and tables will help you to illustrate the findings to your questions?

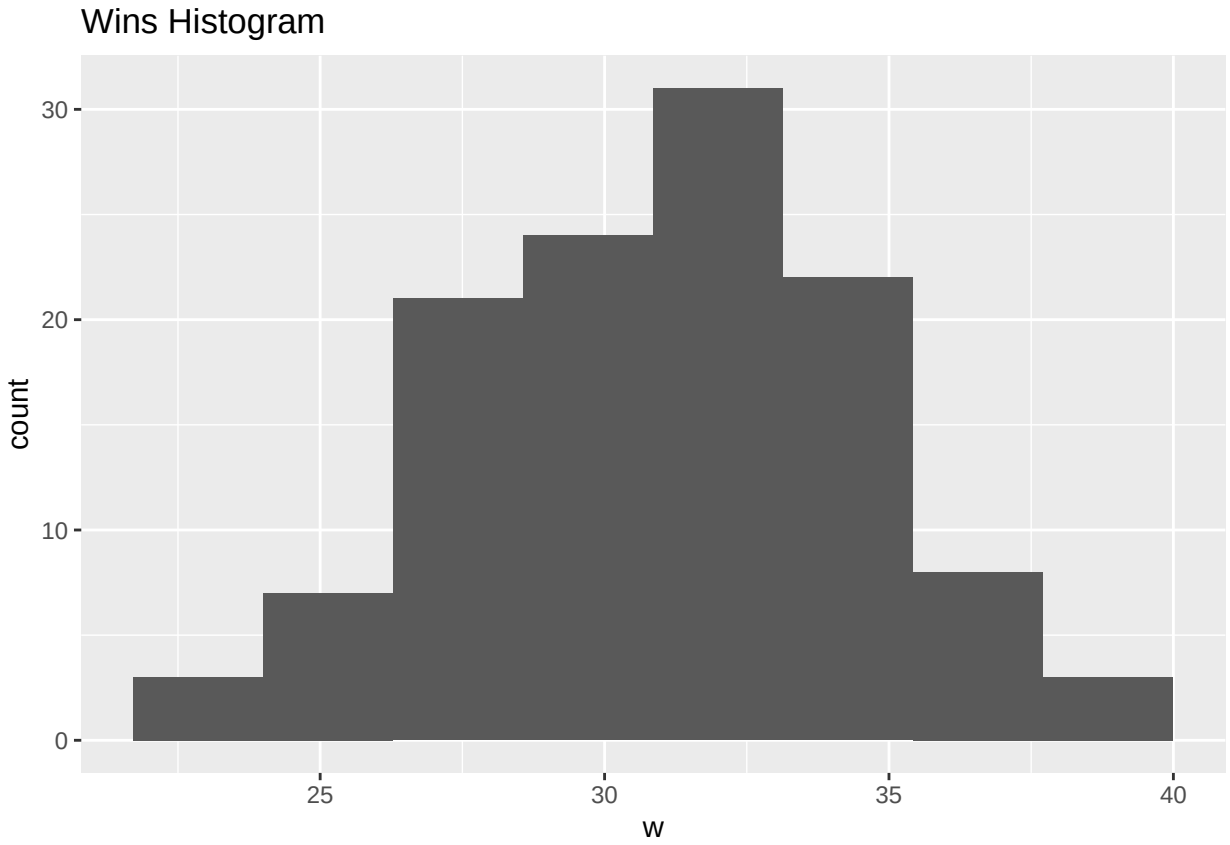
scatterplots, histograms, summary tables

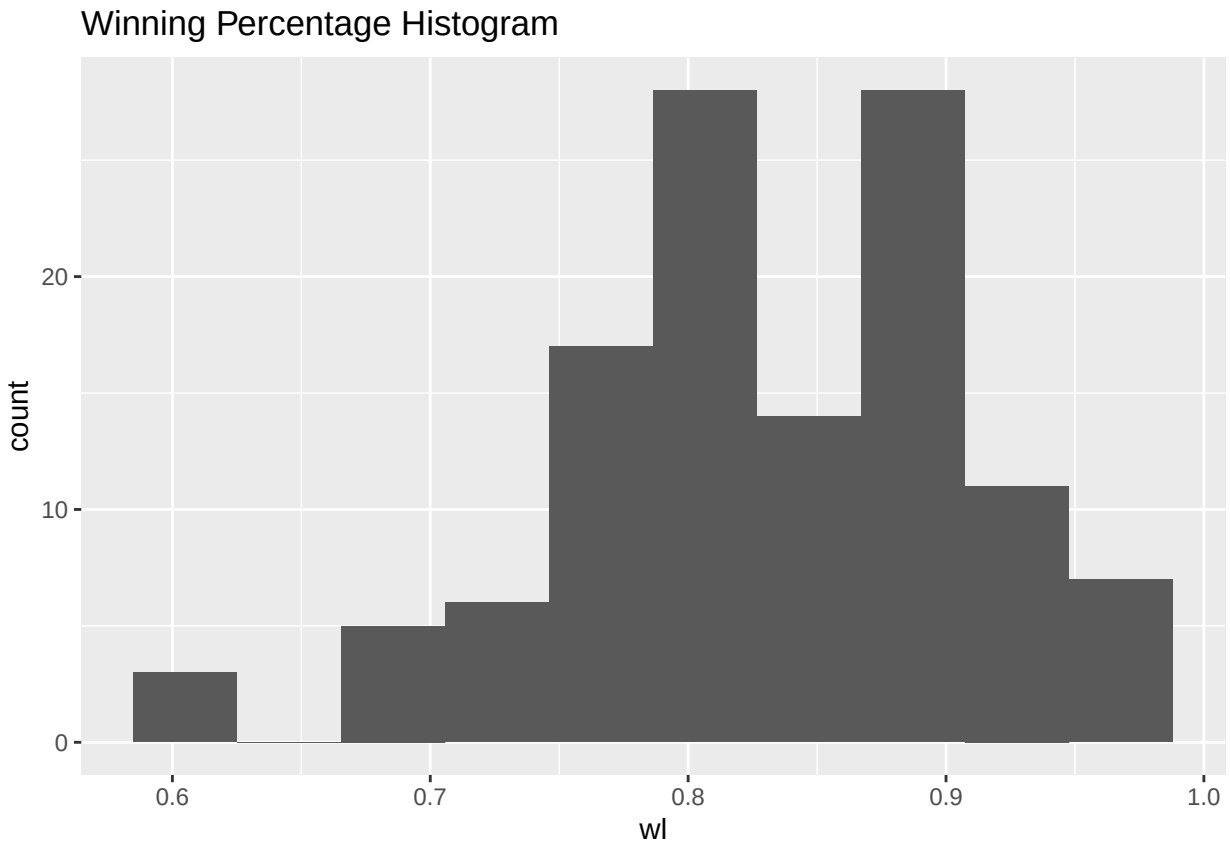

Wins Histogram
count
300
200
100
0
10
20
30
40
w

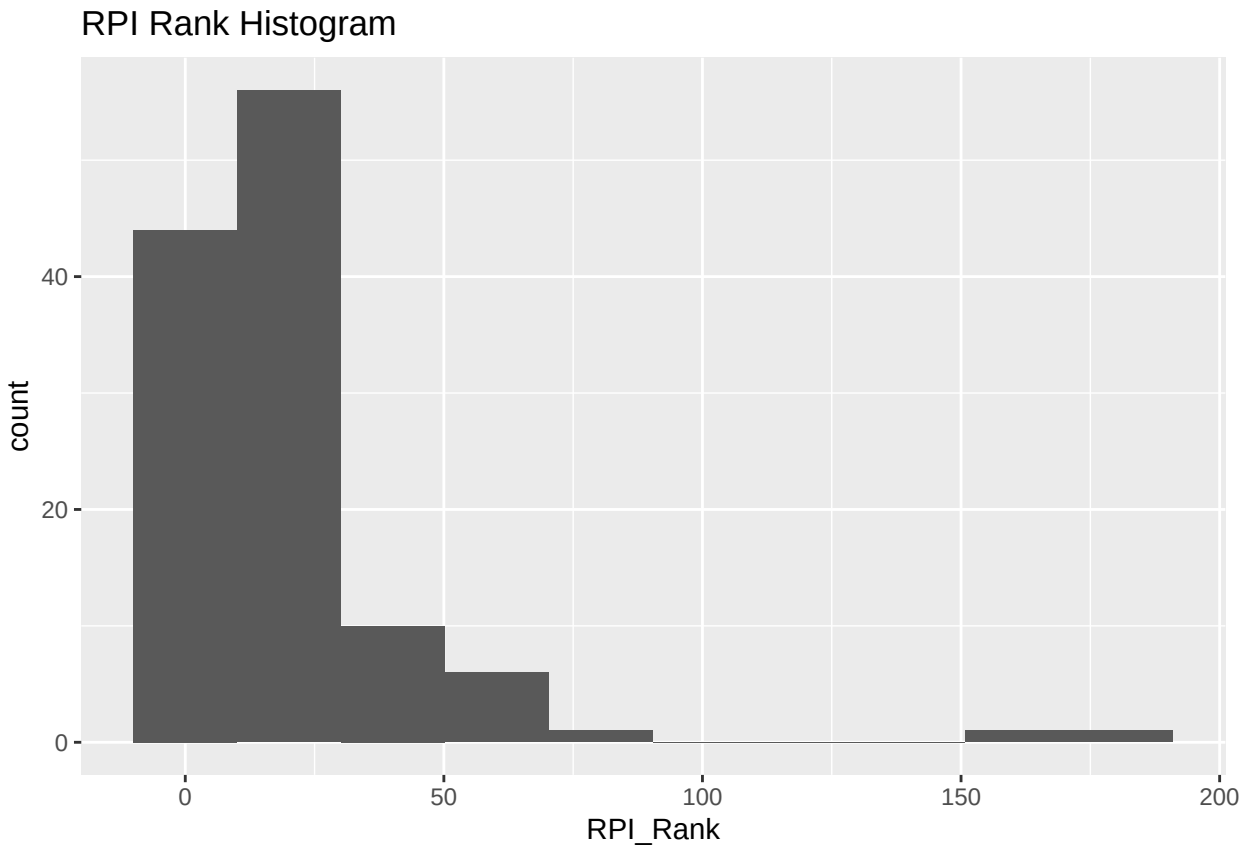

Winning Percentage Histogram
count
200
100
0
0.4
0.6
0.8
1.0
wl

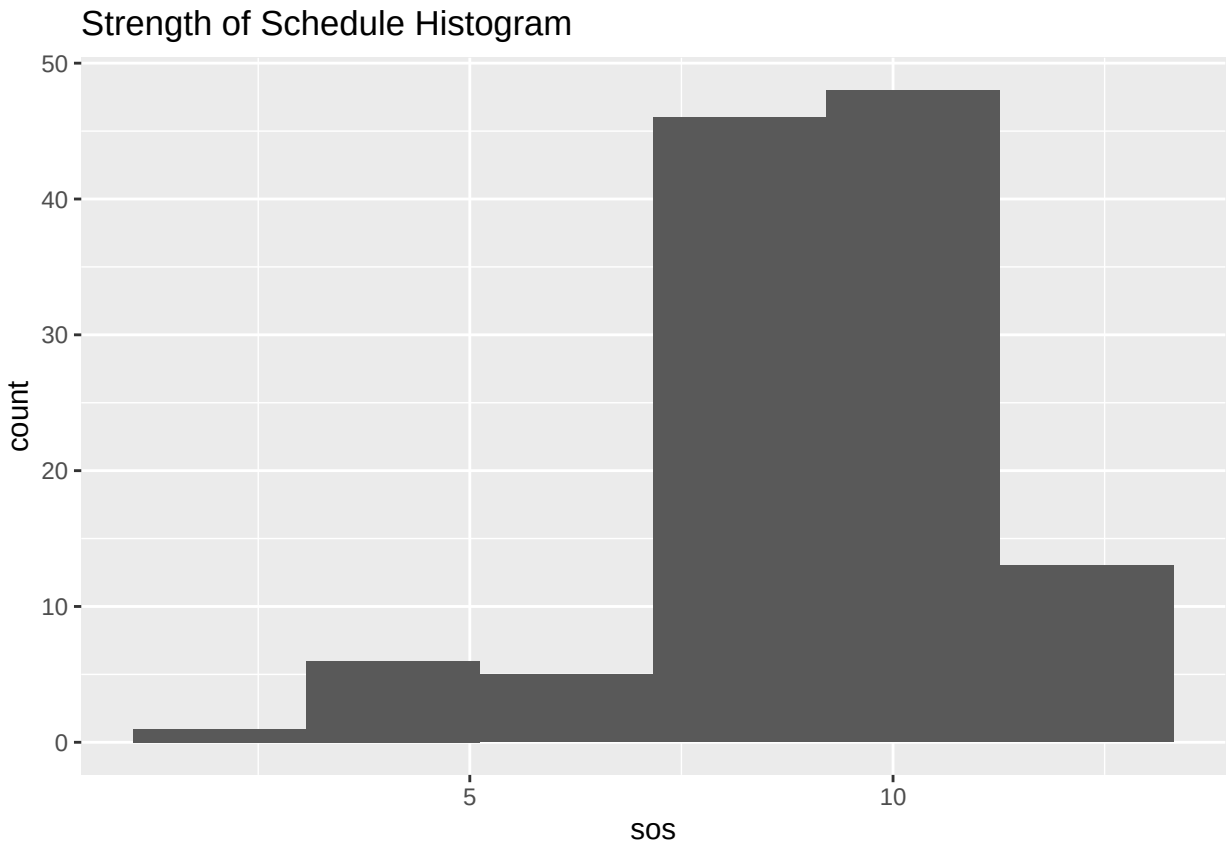

RPI Rank Histogram
count
300
200
100
0
0
100
200
300
RPI_Rank


Strength of Schedule Histogram
count
500
400
300
200
100
0
−10
0
10
sos


Wins Histogram
count
30
20
10
0
25
30
35
40
w


Winning Percentage Histogram
count
20
10
0
0.6
0.7
0.8
0.9
1.0
wl


RPI Rank Histogram
count
40
20
0
0
50
100
150
200
RPI_Rank




```
##          vars   n  mean    sd median trimmed   mad   min    max  range  skew kurtosis   se
## RPI         1 119  0.62  0.03   0.61    0.62  0.03  0.48   0.68   0.21 -0.72     2.20 0.00
## w           2 119 31.02  3.54  31.00   31.04  4.45 22.00  38.00  16.00 -0.09    -0.49 0.32
## RPI_Rank    3 119 19.62 25.18  14.00   14.82 11.86  1.00 182.00 181.00  3.84    19.14 2.31
## sos         4 119  9.20  1.94   9.27    9.35  1.57  2.77  13.01  10.24 -0.81     0.96 0.18
```

### Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I do not plan on incorporation any machine learning techniques at this time to answer my research questions. This may change as I get into the later portions of the course.

# Step 3

## Introduction

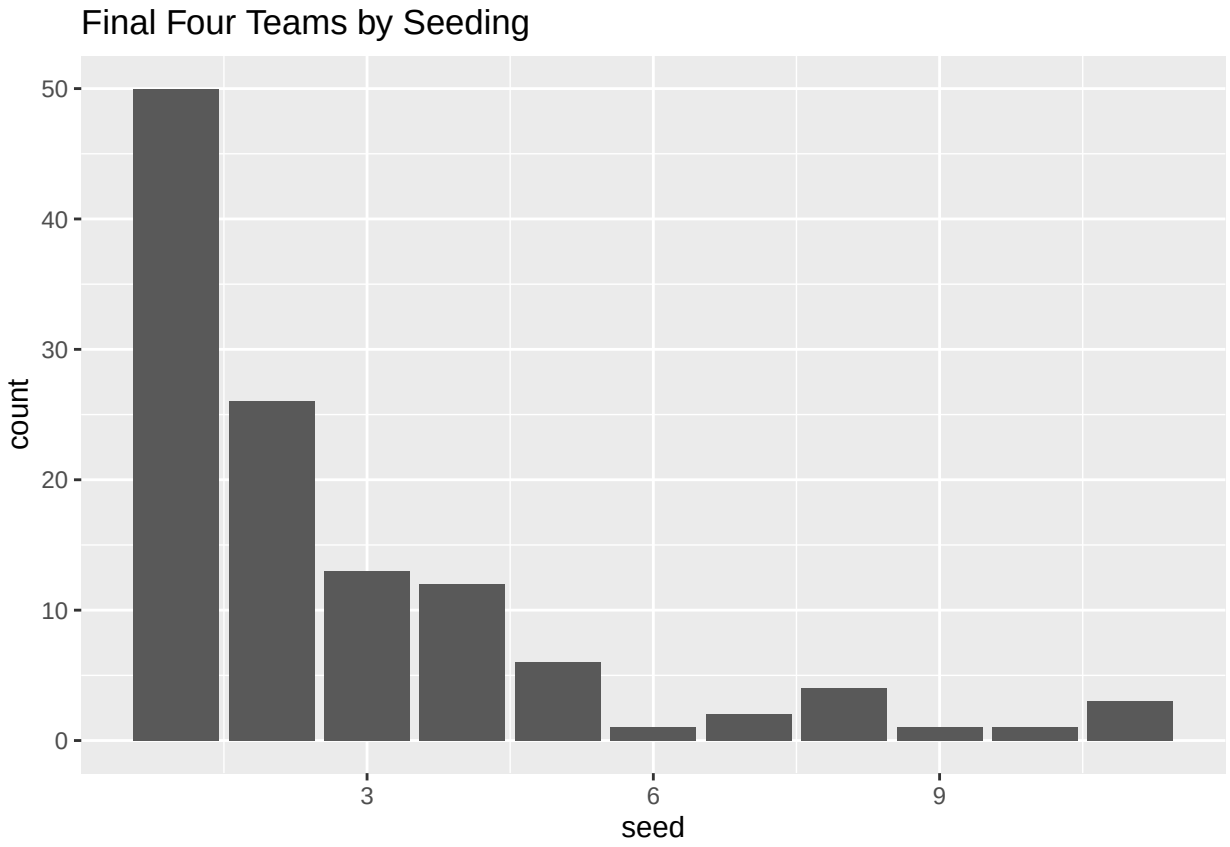
As mentioned above in the Introduction in Step 1, this project is to see if patterns exist that well help predict the Final Four teams of the NCAA March madness tournament.

## Analysis

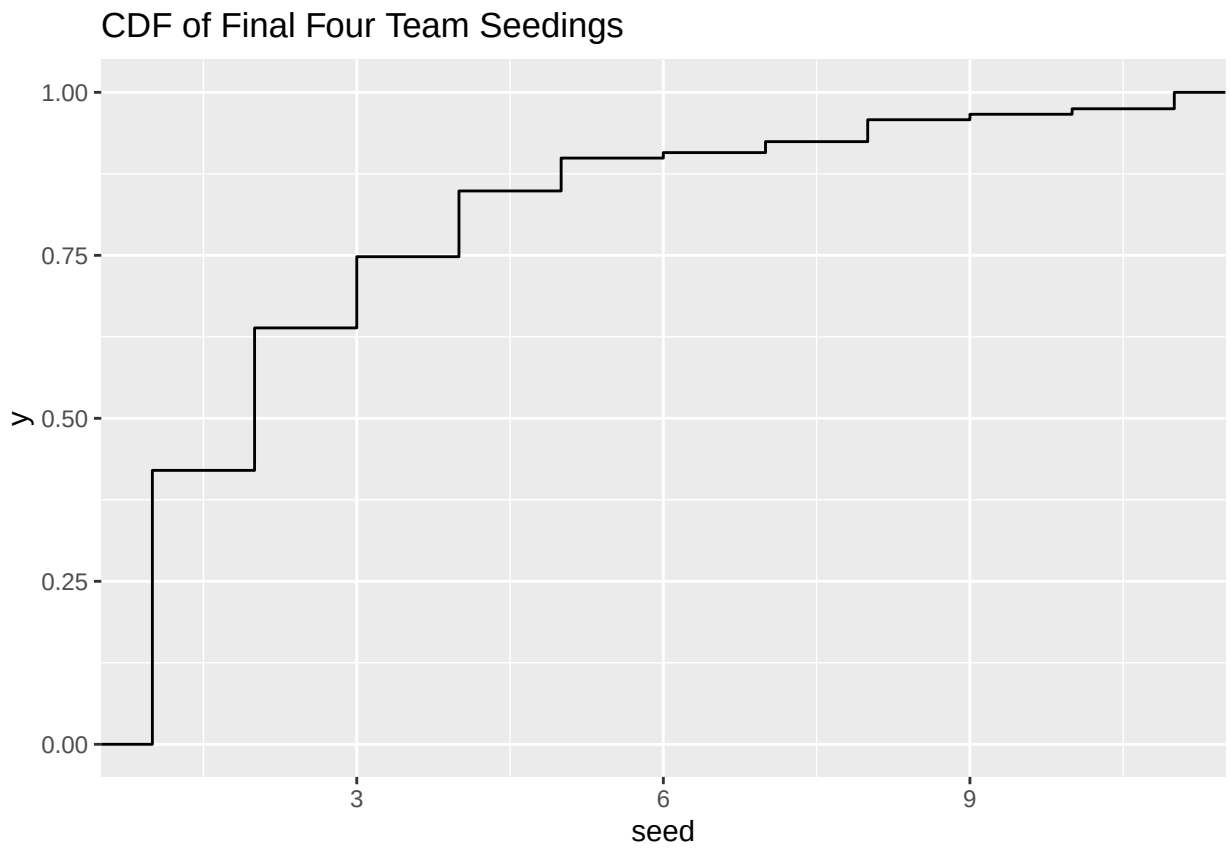
The problem statement is as follows:

**Is there a pattern that exists that will predict the Final Four Teams?**

There seems to be a pattern that exists of being able to predict Final Four teams. I will answer this question by looking at a bar chart of the seedings of teams that have at least made the Final Four.



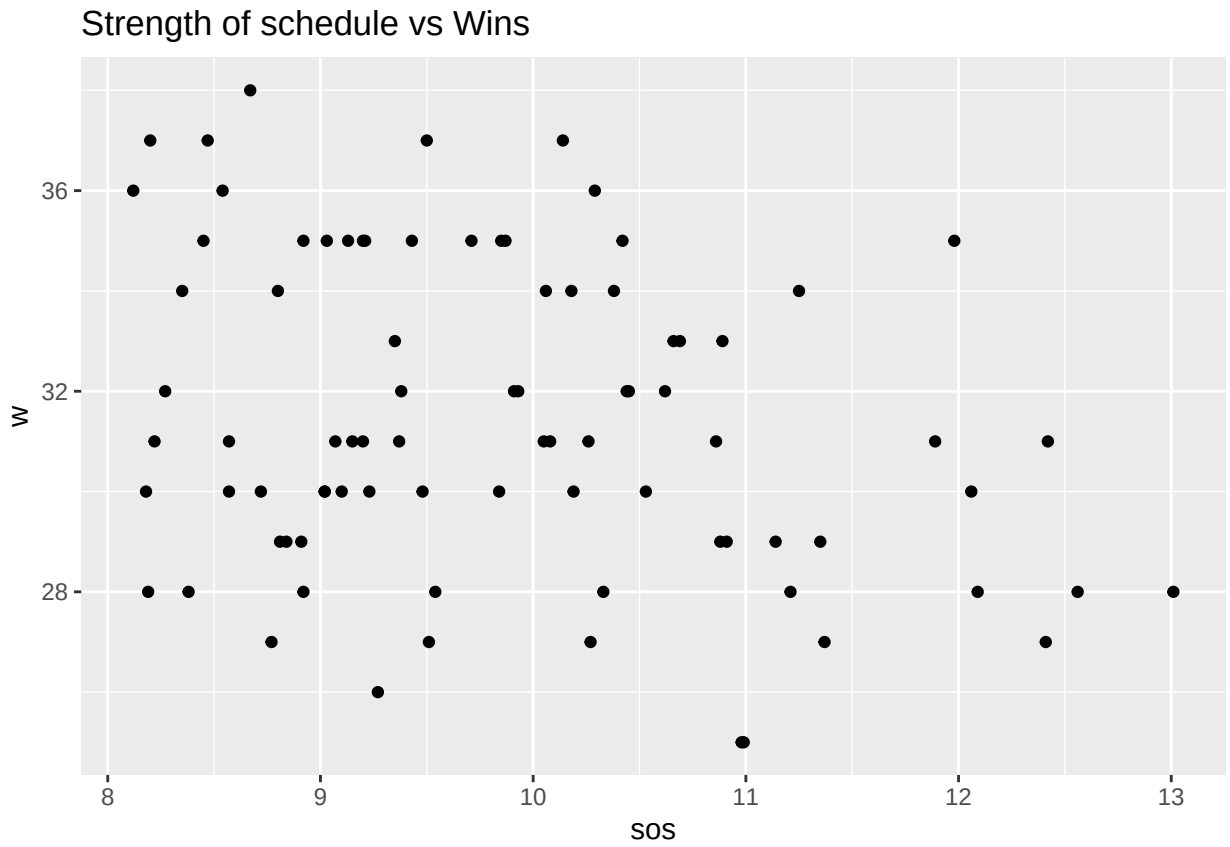

Next, we will look at the cumulative distribution of seedings of Final Four Teams:




This shows that roughly 80% of the Final Four teams are seeded 1-4. This also shows that nearly 50% of the teams that at least make the Final Four are seeded #1.

**What are the most important metrics that will help determine the success of a college basketball team in the tournament?**

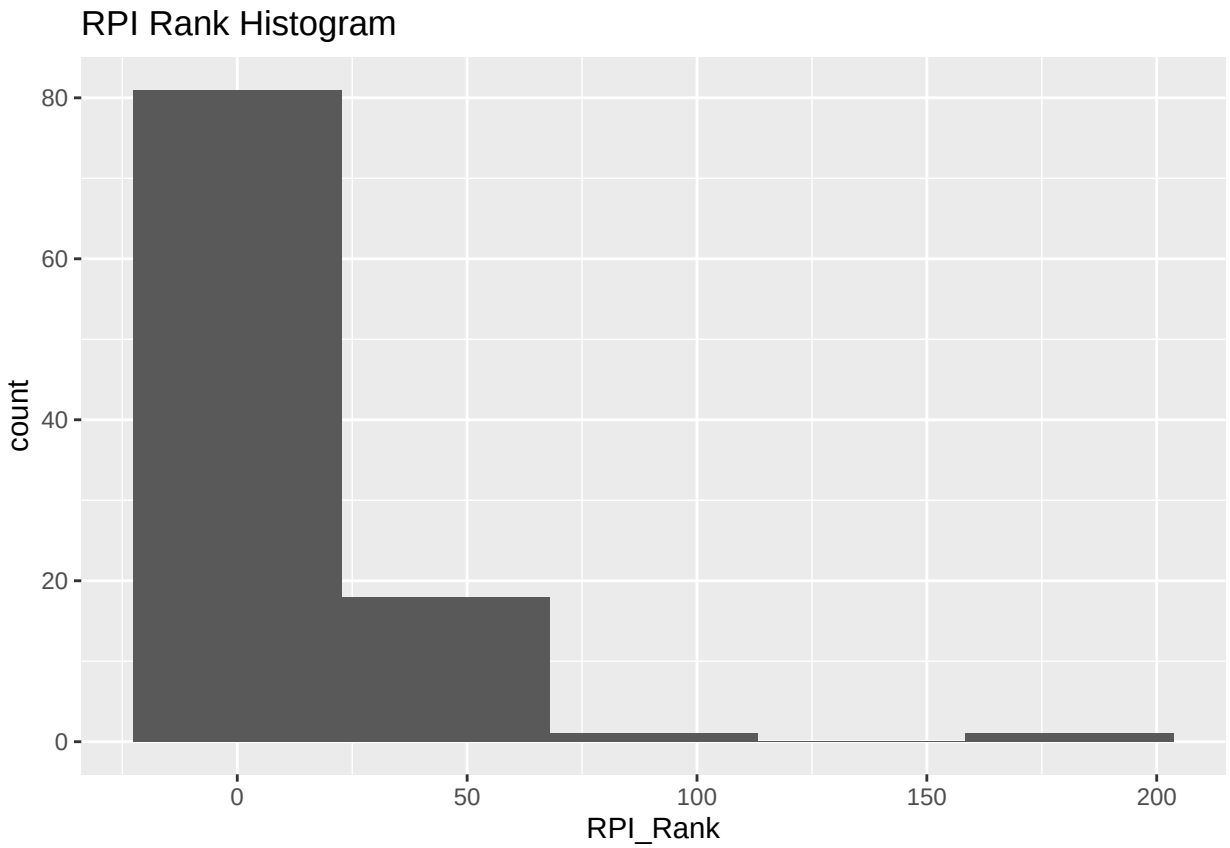
It would appear that winning at least 25 games and having a strength of schedule greater than 9 would be indicators of a team having success in the tournament. We can test this appearance by looking at a scatter plot of #1 seeds who have won at least 25 games with a strength of schedule greater than 9.




```
## [1] -0.2814811
```

After looking at the scatterplot and the corresponding Pearson's correlation coefficient, it would suggest that the more wins you have, the lower your strength of schedule. This would make sense to me because if the team is really good, they are going to beat other really good teams during the regular season (other really good teams = higher strength of schedule), and they are definitely going to beat the not so good teams (not so good teams = lower sos)

**What are some metrics that may not seem important initially but then prove to have statistical significance?**




If your RPI Rank is greater than 25, you could be considered an outlier for making the Final Four. This is something that I did not think would have as much of an effect as it seems to have on predicting the Final Four teams.

**If a pattern exists, will it truly be causation, or is it just a random happening?**

Being a better seeded team and reaching the Final Four is not a random happening. We will look at the teams that made the tournament and get a corrleation between their seeding and their result.

```
## [1] -0.4770564
```

This shows us as the seeding increases, the performance in the NCAA tournament decreases.

If we randomize the seedings against the ncaa_numeric and then get the correlation of the randomized seeds,

```
## [1] 0.006797401
```

it shows us that there the randomized seedings have no correlation to the tournament performance. You can randomize the seedings as many times as you wish, but no significant correlation will come to surface.

**What is the smallest number of predictors that can produce the best model?**

Now we will attempt to produce a model that will help predict the final four teams. The first model will only contain seeding. The second model will include some other metrics, and we will see how the model improves with the additon of those metrics.

```
##
## Call:
```

```
## glm(formula = as.factor(ncaa_numeric) ~ seed, family = binomial(),
##     data = ncaa_data_frame)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1531  -0.8944   0.4553   0.8312   1.8478
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.74598    0.14245   19.28   <2e-16 ***
## seed        -0.26581    0.01435  -18.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2255.5  on 1686  degrees of freedom
## Residual deviance: 1800.9  on 1685  degrees of freedom
## AIC: 1804.9
##
## Number of Fisher Scoring iterations: 4
```

Now for the additional metrics:

```
##
## Call:
## glm(formula = as.factor(ncaa_numeric) ~ seed + w + RPI + srs,
##     family = binomial(), data = ncaa_data_frame)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6095  -0.8440   0.2927   0.7720   2.7364
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.59970    1.21064  -7.103 1.22e-12 ***
## seed        -0.05299    0.03014  -1.758   0.0787 .
## w            0.26548    0.02210  12.011  < 2e-16 ***
## RPI          4.18662    1.65336   2.532   0.0113 *
## srs          0.08129    0.01998   4.068 4.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2255.5  on 1686  degrees of freedom
## Residual deviance: 1595.6  on 1682  degrees of freedom
## AIC: 1605.6
##
## Number of Fisher Scoring iterations: 5
```

There is an improvement from the simple model of seeding when including the predictors wins, RPI and SRS. The residual deviance of the seed only model is 1800 vs 1605 with the additional predictors. One item to note is that by using only wins versus using only seed produces a sligtly better reduction of deviance (1800 vs 1790). It would appear that if we were confined to using only one predictor, wins would be the best choice,

with seeding being a close second.

## Implications

The implications of this project is that hopefully I will be able to better predict the Final Four teams in the future. Based on this analysis, I will look more closely at the seeding a team has, the number of wins they have, and the RPI Ranking for that team.

## Limitations

Limitations that I had due to this project is that I was unable to dig futher into the data as I would have liked. I also wish that I had the time to find the data of the OWP and OOWP columns for the entire data set instead of the years from 2003 on. The most glaring limitation, of course, is the limitation of knowledge that I have regarding my experience with R, and my ability to understand statistics.

## Concluding Remarks

With those limitations being said, I feel that my statistical knowledge has increased greatly since taking this course. I can honestly say that I am beginning to think about the world in a more statistical manner.

I am glad that I chose this topic. It has helped me to gain a better understand of statistics, R, and sports analysis. I look forward to doing more projects like these in the upcoming classes in this curriculum.