Final Project

Determining the Primary Factor in Winning Basketball Games

An Investigative Study By:
Myra Haider
Nikita Hegde
Catherine Hou
Kaushik Ram Ganapathy
Gayatri Mainkar
Chuan Yuan Yeh

INTRODUCTION

The popularity of collegiate sports has been gradually increasing ever since it was introduced. With the establishment of the National Collegiate Athletic Association, or NCAA for short, not only did the rate increase, but commercialization also began to emerge. Consequently, gambling found its way to collegiate sports as grand prizes were given to those who correctly predict the winners of games. One such example is the annual NCAA Women's Division I Basketball Championship. This is a single-game elimination structured tournament where 64 teams enter and one team comes out on top by going undefeated. To correctly predict the winner is very challenging. Thus, it is not surprising that the prizes offered each year are outrageous such as billionaire Warren Buffett's offer of $1 million per year for life for any of his employees who correctly predict the teams that just make it to the Sweet 16[1].

Usually, people went with the team in which they attended, or are attending, for college. However, several statistically-minded ones also attempted to build predictive models based on historical data. Many of the past studies that were done were on predicting match outcomes and comparing different types of machine learning models. This study, however, focuses on the *features* that are important for building a successful model. Specifically, the study investigates whether or not the location of games played can be used as the primary factor for predicting match outcomes due to the *home court advantage* phenomenon. It occurs when teams tend to perform better at their own courts compared to playing at their opposing team's court.

In this study, exploratory data analysis was performed in-depth on historical and current NCAA data to see if there were any anomalies or trends between performance and location. This was followed by hypothesis testings to check if any differences were statistically significant. Finally, a tree model was built to see if the feature *location* was even among the top features used to predict the outcomes. Although the results did not line up with what was speculated in the beginning, this study built a skeleton process which can be used to analyze other features individually to see if they can be used as primary factors to predict match outcomes. This is very important in machine learning nowadays since reducing the features required for a model may prevent overfitting and increase its generalization capability.

---

[1] *March Madness: Warren Buffett offers $1M for life in bracket challenge*. (2019). *usatoday*.

BACKGROUND

NCAA

The NCAA stands for the National Collegiate Athletic Association and is an American organization that legislates and administers intercollegiate athletics[1]. The initial motivation for such an organization was due to an incident back in 1840. During one of the earliest interschool athletic events, it was speculated that Harvard University attempted to gain an advantage over its rival Yale by obtaining the services of an athlete who was not a student[2]. This incident, along with the increasing commercialization of collegiate sports and the numerous injuries, became a call-to-action for multiple university presidents to get together to discuss the potential formation of a regulatory body. With the help from the White House, the Intercollegiate Athletic Association was formed which was renamed as NCAA in 1910.

Due to the rapidly increasing number of higher-level institutions joining the NCAA membership, the organization was split into three divisions in 1973[1]. Each division has its own legislative committees and requirements for institutions to either join or remain as a member. For example, a Division I member institution is required to sponsor at least seven sports for men and seven for women with two team sports for each gender while a Division II member only requires five[3]. There is a common misconception that only Division I schools have the best athletes. However, the purpose of splitting the NCAA into three divisions is to also group like-minded campuses in the areas of philosophy and opportunity, and not just for the sake of competition.

As time passed, the NCAA expanded its functions from merely formulating and enforcing rules of plays for various sports to include television coverage. Furthermore, it also began conducting national championship events in 1921 until today.

NCAA WOMEN'S DIVISION I BASKETBALL

Women's basketball was introduced to the NCAA championship program in the 1981-82 school year[4]. Although the national championship, also known as March Madness, starts in mid-March, the women's basketball season officially begins in early November of each year. For six months, thousands of games are played between the Division I schools as each team fights

---

[1] Encyclopedia Britannica. (2019). *National Collegiate Athletic Association | American organization*. [online] Available at: https://www.britannica.com/topic/National-Collegiate-Athletic-Association [Accessed 13 Mar. 2019].
[2] Rodney K. Smith, The National Collegiate Athletic Association's Death Penalty: How Educators Punish Themselves and Others, 62 IND. L.J. 985, 988-89 (1987) [hereinafter Smith, Death Penalty]; Rodney K. Smith, Little Ado About Something: Playing Games With the Reform of Big-Time Athletics, 20 CAP. U. L. Rev. 567, 569-70 (1991) [hereinafter Smith, Little Ado].
[3] *Divisional Differences and the History of Multidivision Classification*. (2013). *NCAA.org - The Official Site of the NCAA*. Retrieved 14 March 2019, from http://www.ncaa.org/about/who-we-are/membership/divisional-differences-and-history-multidivision-classification
[4] *DI Women's Basketball Championship History | NCAA.com*. (2019). *Ncaa.com*.

for a spot in the championship tournament. Since 2014, the NCAA basketball season is formatted so that only a total of 64 teams can qualify for the tournament played in March. Out of the 64 teams, 32 of those are entered as *automatic bids* and the remaining teams are selected as *at-large bids*.

A team may earn an *automatic bid* by winning its respective conference tournament. The conferences are subdivisions within the Division I organization which include the Ivy, Pac-12, ACC, and several others. Unlike the procedure of splitting the NCAA into three divisions, the conferences are split based on geography and school systems. As an illustration, the Big West conference comprises of institutions situated along the West coast while the Ivy conference consists of only the Ivy League schools[5]. Teams that do not win their conference tournaments still have a shot at the national championship through *at-large bids*. This relies on the NCAA Selection Committee to select the remaining teams on Monday (more commonly known as *Selection Monday*) following the last regular season game. The selection process is based on numerous factors including, but not limited to, game results, strength of schedules, location, and scoring margin[6]. Another process called *seeding* occurs after all 64 teams have been confirmed. The 64 teams are ranked based on similar features from the selection process and then 16 groups of four. Group 1, which consists of teams ranked #1 to #4, are the #1 seeds, Group 2 are the #2 seeds, and so on. Each team is then split into four regions according to their geography so that there are four regions of 16 teams each ranked from 1 through 16. The final procedure is to draw the tournament bracket for the first round such that the top seed plays the bottom seed, the second seed plays the second to last seed, and so forth (refer to Figure 1). The tournament that begins in a single-elimination fashion until one team wins it all.

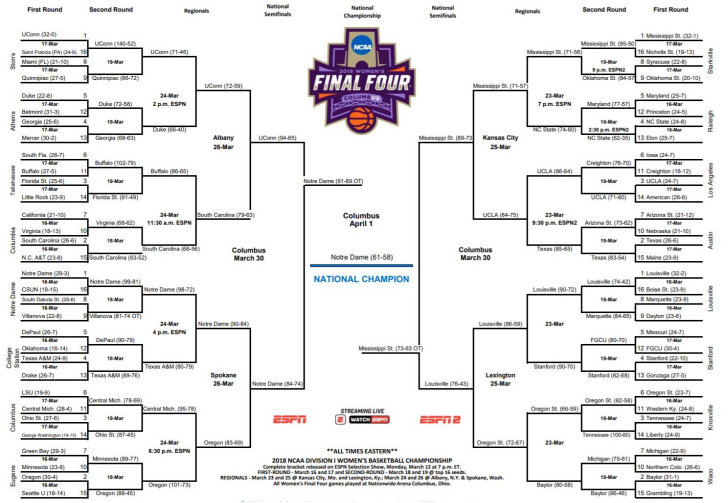


Figure 1. 2018 NCAA Division I Women's Basketball Championship Bracket[7]

[5] GamesCricketRugbyCFL, &., League, N., Softball, N., Sports, O., FB, R., & BB, R. et al. (2019). *Pac-12 Conference Standings - Women's College Basketball - ESPN. ESPN.com*.

[6] GamesCricketRugbyCFL, &., League, N., Softball, N., Sports, O., FB, R., & BB, R. et al. (2019). *Nothing but NET: NCAA boots RPI for evaluation. ESPN.com*.

[7] *2018 Division I Women's Basketball Official Bracket | NCAA.com*. (2019). *Ncaa.com*.

Data

The dataset from the Kaggle website is comprehensive and contains a large number of datasets. Each dataset contains numerous statistical variables, each with a definitive structure which is reflective of the highly structured format followed in gathering information on NCAA Basketball games. From a statistical standpoint, the kind of variables recorded range from the nominal variables such as Team ID, to ranging to highly quantitative variables such as the seed number (*indicative of the ranking)* of the particular team. In short, a short list of the data collected is shown below:

A. Team ID with the corresponding Team Names.
B. Tournament seeds *(Rankings)* since the 1997-98 NCAA season.
C. Final scores of all regular season/s, conference tournament/s, and NCAA tournament games since 1997-98 season.
D. Season-level details including dates and region names.

As we can see, the data contains various kinds of statistical variables being measured. For instance, in particular, during analysis details such as the date of various matches quantify when the match happen in time, and special attention must be paid in handling data-typed which deal with time.

SEASONS

Owing to the highly structured form of the data, special attention needs to be paid regarding to the season numbers. However, first, it is important to understand the timelines of NCAA basketball seasons.

It is a well known fact that the college basketball season starts from early November, and lasts till the Middle of March. This is known as the "regular-college basketball season." After this, the top 64 teams are selected, *(by their performance during the regular season)*, to proceed to the National Championship or the "tournament. The tournament usually starts from the middle of March. However, this start-date can be as late as the first week of April, and varies from one season to another. Thus, we can see that the totality of each season spans two calendar years.

To avoid any ambiguities in identifying the correct year/s in which a season took place, this dataset, uses the convention of using the year in which the season end on. This argument is logically sound since the championship round of the NCAA tournament does take place in the year in which the season ends on and thus, identifying the season by the year in which it ends indirectly creates a mapping between the season number, and the year in which the championship was played! However, this convention in not universal and care must be taken when factoring in external information such as sources and/or additional datasets in the analysis.

The Kaggle Dataset provides a useful example to explain the same. The documentation in the dataset claims that the current season will be identified in our data as the 2019 season, not the 2018 season or the 2018-19 season or the 2018-2019 season, though you may see any of these in everyday use outside of our data.

DIVISIONS

In the NCAA, the various schools across the U.S are organized into divisions. The NCAA sets guidelines for the classification of schools into one of 3 divisions. These divisions are labelled as Divisions I, Division II, and Division III Respectively. The description between all of these divisions is described below:

DIVISION-I

The Division I schools are the premier teams in the NCAA. They have the biggest student bodies, huge amounts of budget for athletics and offer extensive scholarships for students. There are currently around 353 Division I schools in the NCAA. An interesting fact is that all of the Ivy League schools are Division I but offer no athletic scholarships. There are specific rules set for specific sports as well. For instance, Men's and Women's Basketball teams MUST play all but two of their games against Division I schools. Further, each playing season has to be represented by each of the genders. More rules and specifications can be found on the NCAA website.

DIVISION-II

The Division II Schools are extremely similar to Division I, except that the amount of athletic scholarships they give-out is far lower. Currently there are about 300 Division II Schools scattered across the U.S. While a typical NCAA Athlete at a Division I school enjoys a full academic scholarship, the athletes at Division II schools, most of the time, receive only partial athletic scholarships. Further, Division II schools also tend to travel within their own region rather than nationally like Division I Schools. Division II Basketball teams must play at least 50% of their games with Division II schools. Further unlike Division I schools, there are no pre-defined attendance requirements for Division II Schools.

DIVISION-III

The Division III Schools are the most common school type across the U.S. Currently there are about 444 Division III Schools in the U.S, comprising of more than 170,000 players. A key difference between Division III and Division I and Division II is that Athletes at Division III schools receive no Athletic Scholarships. However, a majority of the athletes are on some form of academic or need-based aid. Further unlike the Division I and Division II Schools, Division II has no set attendance, contest or participation through the season.

FILE FORMAT OF THE DATA

The data collected from the Kaggle website is organised in the form of a comma separated (.csv) files, which proves to be useful since the dataset can then be plugged directly into existing pipelines such as Pandas in the Python Programming Language, which can then be used directly for analysis. Extensive tests were made to verify that the data was indeed un-corrupted and followed the specifications to which a .csv file must be formatted.

Now, we proceed to describe each file, along with any underlying special features about the structure of the data contained within these files, followed by a detailed description of what each variable in the dataset represents. The materials for variable description are directly sourced from the Kaggle website when present, and are otherwise written in a format comparable to what the dataset has natively used. The file name in which the data exists is the header of the section in bold. By default the end of the file as specified in the section above is a .csv. However, if a special formatting does exist it is indicated by an explicit statement on the first line of the paragraph following the header before attempting any description of the variables in the dataset.

File: WTeams

Format: *csv*

Description:

This file identifies the different college teams present in the dataset. Each school is uniquely identified by a 4 digit id number. The Games are listed only when the two teams are Divison-I teams. This must be factored when performing any analysis! There are 353 teams currently in Division-I, and an overall total of 366 teams in the team listing (*each year, some teams might start being Division-I programs, and others might stop being Division-I programs)*.

This year there are two teams that are new to Division I: Cal Baptist (TeamID=3465) and North Alabama (TeamID=3466), and any historical data for these teams prior to the current season is by-definition absent!

Fields:

1. TeamID - a 4 digit id number, from 3000-3999, uniquely identifying each NCAA Women's Team. This field can uniquely identify a team across seasons since it does not vary from season to season. To avoid the possible confusion between the men's data and the women's data, all of the men's team ID's range from 1000-1999, whereas all of the women's team ID's range from 3000-3999.

2. TeamName - a compact spelling of the team's college name, 16 characters or fewer. There are no commas or double-quotes in the team names, but you will see some characters that are not letters or spaces, such as “&”, “ ’ ”, and “-” .

File: WSeasons

Format: *csv*

Description:

This file identifies the different seasons included in the historical data, along with certain season-level properties.

Fields:

1. Season - indicates the year in which the tournament was played. Care must be taken to correctly interpret the Season number by following the guidelines in the writeup on Seasons above!

2. DayZero - The date corresponding to day-num=0 during that season. All game dates have been aligned upon a common scale so that in each season Selection Monday is on day 133. All game data includes the day number in order to make it easier to perform date calculations. If you need to know the exact date a game was played on, you can combine the game's "day-num" with the season's "day-zero"!

3. RegionW, RegionX, Region Y, Region Z - by convention, the four regions in the final tournament are always named W, X, Y, and Z. Whichever region's name comes first alphabetically, that region will be Region W. And whichever Region plays against Region W in the national semifinals, that will be Region X. For the other two regions, whichever region's name comes first alphabetically, that region will be Region Y, and the other will be Region Z. The final W/X/Y/Z designations is unknown until Selection Monday, because the national semifinal pairings in the Final Four will depend upon the overall ranks of the four #1 seeds.

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 also known as Selection Monday. 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.

File: WNCAATourneySeeds

Format: *csv*

Description:

This file identifies the seeds for all teams in each NCAA tournament, for all seasons of historical data. Thus, there are exactly 64 rows for each year, since there are no play-in teams in the women's tournament. We will not know the seeds of the respective tournament teams, or even exactly which 64 teams it will be, until Selection Monday on March 18, 2019.

Fields:

A. Season - the year that the tournament was played in.

B. Seed - this is a 3-character identifier of the seed, where the first character is either W, X, Y, or Z (identifying the region the team was in) and the next two digits is the seed within the region.

C. TeamID - this identifies the id number of the team, as specified in the WTeams.csv file

---

File: WRegularSeasonDetailedResults

Format: *csv*

Description:

This file identifies the game-by-game results for many seasons of historical data, starting with the 2010 season. For each season, the file includes all games played from daynum 0 through 132.

It is important to realize that the "Regular Season" games are simply defined to be all games played on DayNum=132 or earlier (DayNum=133 is Selection Monday). Thus a game played before Selection Monday will show up here whether it was a pre-season tournament, a non-conference game, a regular conference game, a conference tournament game, or any other game!

It also has the the box score which is used to summarize/average the data of Games played (GP), Games started (GS), Minutes Played (MIN or MPG), Field-goals made (FGM), Field-goals attempted (FGA), Field goal percentage (FG%), 3-pointers made (3PM), 3-pointers attempted (3PA), 3-point field goal (3P%), Free throws made (FTM), Free throws attempted (FTA), Free

throw percentage (FT%), Offensive Rebounds (OREB), Defensive Rebounds (DREB), Total rebounds (REB), Assists (AST), Turnovers (TOV), Steals (STL), Blocked shots (BLK), Personal fouls (PF), Points scored (PTS), and Plus/Minus for Player efficiency (+/-)

Fields:

1. Season - this is the year of the associated entry in WSeasons.csv (the year in which the final tournament occurs).

2. DayNum - this integer always ranges from 0 to 132, and tells you what day the game was played on. It represents an offset from the "DayZero" date in the "WSeasons.csv" file. There are no teams that ever played more than one game on a given date!

3. WTeamID - this identifies the id number of the team that won the game, as listed in the "WTeams.csv" file. No matter whether the game was won by the home team or visiting team, or if it was a neutral-site game, the "WTeamID" always identifies the winning team.

4. WScore - this identifies the number of points scored by the winning team.

5. LTeamID - this identifies the id number of the team that lost the game.

6. LScore - this identifies the number of points scored by the losing team. Thus you can be confident that WScore will be greater than LScore for all games listed.

7. NumOT - this indicates the number of overtime periods in the game, an integer 0 or higher.

8. WLoc - this identifies the "location" of the winning team. If the winning team was the home team, this value will be "H". If the winning team was the visiting team, this value will be "A". If it was played on a neutral court, then this value will be "N".

9. WFGM - field goals made (by the winning team)

10. WFGA - field goals attempted (by the winning team)

11. WFGM3 - three pointers made (by the winning team)

12. WFGA3 - three pointers attempted (by the winning team)

13. WFTM - free throws made (by the winning team)

14. WFTA - free throws attempted (by the winning team)

1. WOR - offensive rebounds (pulled by the winning team)

15. WDR - defensive rebounds (pulled by the winning team)

16. WAst - assists (by the winning team)

17. WTO - turnovers committed (by the winning team)

18. WStl - steals (accomplished by the winning team)

19. WBlk - blocks (accomplished by the winning team)

20. WPF - personal fouls committed (by the winning team)

21. LFGM - field goals made (by the winning team)

22. LFGA - field goals attempted (by the winning team)

23. LFGM3 - three pointers made (by the winning team)

24. LFGA3 - three pointers attempted (by the winning team)

25. LFTM - free throws made (by the winning team)

26. LFTA - free throws attempted (by the winning team)

27. LOR - offensive rebounds (pulled by the winning team)

28. LDR - defensive rebounds (pulled by the winning team)

29. LAst - assists (by the winning team)

30. LTO - turnovers committed (by the winning team)

31. LStl - steals (accomplished by the winning team)

32. LBlk - blocks (accomplished by the winning team)

33. LPF - personal fouls committed (by the winning team)

By convention, "field goals made" *(either WFGM or LFGM)* refers to the total number of fields goals made by a team, a combination of both two-point field goals and three-point field goals. And "three point field goals made" *(either WFGM3 or LFGM3)* is just the three-point fields goals made. To calculate two-point field goals, subtracting WFGM from WFGM3 will yield the required result. And the total number of points scored is most simply expressed as 2*FGM + FGM3 + FTM.

File: WNCAATourneyDetailedResults

Format: *csv*

Description:

This file identifies the game-by-game NCAA tournament results for all seasons of historical data. The data is formatted exactly like the WRegularSeasonDetailedResults.csv data. Each season you will see 63 games listed, since there are no women's play-in games.

It also has the the box score which is used to summarize/average the data of Games played (GP), Games started (GS), Minutes Played (MIN or MPG), Field-goals made (FGM), Field-goals attempted (FGA), Field goal percentage (FG%), 3-pointers made (3PM), 3-pointers attempted (3PA), 3-point field goal (3P%), Free throws made (FTM), Free throws attempted (FTA), Free throw percentage (FT%), Offensive Rebounds (OREB), Defensive Rebounds (DREB), Total rebounds (REB), Assists (AST), Turnovers (TOV), Steals (STL), Blocked shots (BLK), Personal fouls (PF), Points scored (PTS), and Plus/Minus for Player efficiency (+/-)

NOTE ON THE DAYS OF MATCHES

There have been four different schedules over the course of the past 20 years for the women's tournament, as follows:

2017-2019 Season

A. Round 1 = days 137/138 (Fri/Sat)
B. Round 2 = days 139/140 (Sun/Mon)
C. Round 3 = days 144/145 (Sweet Sixteen, Fri/Sat)
D. Round 4 = days 146/147 (Elite Eight, Sun/Mon)
E. National Semi-Final = day 151 (Fri)
F. National Final = day 153 (Sun)

2015-2016 Season

A. Round 1 = days 137/138 (Fri/Sat)
B. Round 2 = days 139/140 (Sun/Mon)
C. Round 3 = days 144/145 (Sweet Sixteen, Fri/Sat)
D. Round 4 = days 146/147 (Elite Eight, Sun/Mon)
E. National Semi-Final = day 153 (Sun)
F. National Final = day 155 (Tue)

2003-2014 Season

A. Round 1 = days 138/139 (Sat/Sun)
B. Round 2 = days 140/141 (Mon/Tue)
C. Round 3 = days 145/146 (Sweet Sixteen, Sat/Sun)
D. Round 4 = days 147/148 (Elite Eight, Mon/Tue)
E. National Semi-Final = day 153 (Sun)
F. National Final = day 155 (Tue)

1998-2002 Season

A. Round 1 = days 137/138 (Fri/Sat)
B. Round 2 = days 139/140 (Sun/Mon)
C. Round 3 = day 145 only (Sweet Sixteen, Sat)
D. Round 4 = day 147 only (Elite Eight, Mon)
E. National Semi-Final = day 151 (Fri)
F. National Final = day 153 (Sun)

Fields:

1. Season - this is the year of the associated entry in WSeasons.csv (the year in which the final tournament occurs).

2. DayNum - this integer always ranges from the start date from the tables corresponding to the season above to the national final date as per the values given corresponding to the respective season above, and tells you what day the game was played on. It represents an offset from the "DayZero" date in the "WSeasons.csv" file. There are no teams that ever played more than one game on a given date!

3. WTeamID - this identifies the id number of the team that won the game, as listed in the “WTeams.csv" file. No matter whether the game was won by the home team or visiting team, or if it was a neutral-site game, the "WTeamID" always identifies the winning team.

4. WScore - this identifies the number of points scored by the winning team.

5. LTeamID - this identifies the id number of the team that lost the game.

6. LScore - this identifies the number of points scored by the losing team. Thus you can be confident that WScore will be greater than LScore for all games listed.

7. NumOT - this indicates the number of overtime periods in the game, an integer 0 or higher.

8. WLoc - this identifies the "location" of the winning team. If the winning team was the home team, this value will be "H". If the winning team was the visiting team, this value will be "A". If it was played on a neutral court, then this value will be “N".

9. WFGM - field goals made (by the winning team)

10. WFGA - field goals attempted (by the winning team)

11. WFGM3 - three pointers made (by the winning team)

12. WFGA3 - three pointers attempted (by the winning team)

13. WFTM - free throws made (by the winning team)

14. WFTA - free throws attempted (by the winning team)

15. WOR - offensive rebounds (pulled by the winning team)

16. WDR - defensive rebounds (pulled by the winning team)
17. WAst - assists (by the winning team)
18. WTO - turnovers committed (by the winning team)
19. WStl - steals (accomplished by the winning team)
20. WBlk - blocks (accomplished by the winning team)
21. WPF - personal fouls committed (by the winning team)
22. LFGM - field goals made (by the winning team)
23. LFGA - field goals attempted (by the winning team)
24. LFGM3 - three pointers made (by the winning team)
25. LFGA3 - three pointers attempted (by the winning team)
26. LFTM - free throws made (by the winning team)
27. LFTA - free throws attempted (by the winning team)
28. LOR - offensive rebounds (pulled by the winning team)
29. LDR - defensive rebounds (pulled by the winning team)
30. LAst - assists (by the winning team)
31. LTO - turnovers committed (by the winning team)
32. LStl - steals (accomplished by the winning team)
33. LBlk - blocks (accomplished by the winning team)
34. LPF - personal fouls committed (by the winning team)

The datasets described below are pertaining to the geographic distribution of games, with the respective data recorded by location. The data is recorded for both, regular and tournament games since the 2010 season.

Further, external datasets were also collected to supplement the existing datasets during analysis, and the formats of these external datasets vary considerably to those from the NCAA dataset. Hence, code was written to post-process them in a format similar to that of the one used by the NCAA datasets.

File: WCities

Format: *csv*

Description:

This file provides a master list of cities that have been locations for games played, ever since the 2010 season.

Fields:

1. CityID - a four-digit ID number uniquely identifying a city.

2. City - the text name of the city.

3. State - the state abbreviation of the state that the city is in. In a few rare cases, the game location is not inside one of the 50 U.S. states and so other abbreviations are used, for instance Cancun, Mexico has a state abbreviation of MX.

File: WGameCities

Format: *csv*

Description:

This file identifies all games, starting with the 2010 season, along with the city that the game was played in. Games from the regular season and the NCAA® tourney are all listed together.

Fields:

1. Season - this is the year of the associated entry in WSeasons.csv (the year in which the final tournament occurs).

2. DayNum - this integer always ranges from the start date from the tables corresponding to the season above to the national final date as per the values given corresponding to the respective season above, and tells you what day the game was played on. It represents an offset from the "DayZero" date in the "WSeasons.csv" file. There are no teams that ever played more than one game on a given date!

3. WTeamID - this identifies the id number of the team that won the game, as listed in the “WTeams.csv" file. No matter whether the game was won by the home team or visiting team, or if it was a neutral-site game, the "WTeamID" always identifies the winning team.

   LTeamID - this identifies the id number of the team that lost the game.

4. CRType - this can be either Regular or NCAA. If it is Regular, you can find more about the game in the WRegularSeasonCompactResults.csv file. If it is NCAA, you can find more about the game in the WNCAATourneyCompactResults.csv file.

5. CityID - the ID of the city where the game was played, as specified by the CityID column in the WCities.csv file.

File: Cities

Format: *csv*

Description:

This file identifies all the cities in the Unites States along with their States, Counties, Latitudes and Longitudes. The original dataset contained many irrelevant parameters such as population, density, timezone amongst others. These were dropped from the dataset during analysis and hence detailed descriptions are not included for these un-used parameters.

Fields:

1. City - This is the name of the City in the United States.

2. State_id - This is a two letter abbreviation for the state. For instance, Washington State will be WA and California will be CA etc. for each city

3. Lat-This is the latitude of the city in the U.S

4. Lng- This is the Longitude of the city in the U.S

File: States

Format: *csv*

Description:

This file identifies all the States in the Unites States along with their region as classified by the U.S Census Bureau. The region is the classification region such as Midwest, South, North-East or West corresponding to the geographical location of the state within the US. The original dataset contained irrelevant parameters such as the Division and these were dropped from the dataset during analysis and hence detailed descriptions are not included for these un-used parameters.

Fields:

1. State - This is the name of the State in the United States.

2. State Code - This is a two letter abbreviation for the state. For instance, Washington State will be WA and California will be CA etc. for each city

3. Region - This is the region to which a particular U.S State belongs to The value of Region may be Midwest, West, North-East or South depending upon the geographical location of the state within the U.S.

File: States.geojson

Format: *geojson*

Description:

This file creates a geojson file along, which ultimately encodes all the states within the states along with the respective geographic boundaries within these states. This is accomplished by using Polygon objects within the json file. The arrays of coordinates correspond to the vertices of the polygon which in turn corresponds to lines segments on the boundaries of the various states. The boundaries of each one of these states are accessible through an id parameter.

Visualization:

Here is the raw visualization of the used geojson file rendered in a web-browser.

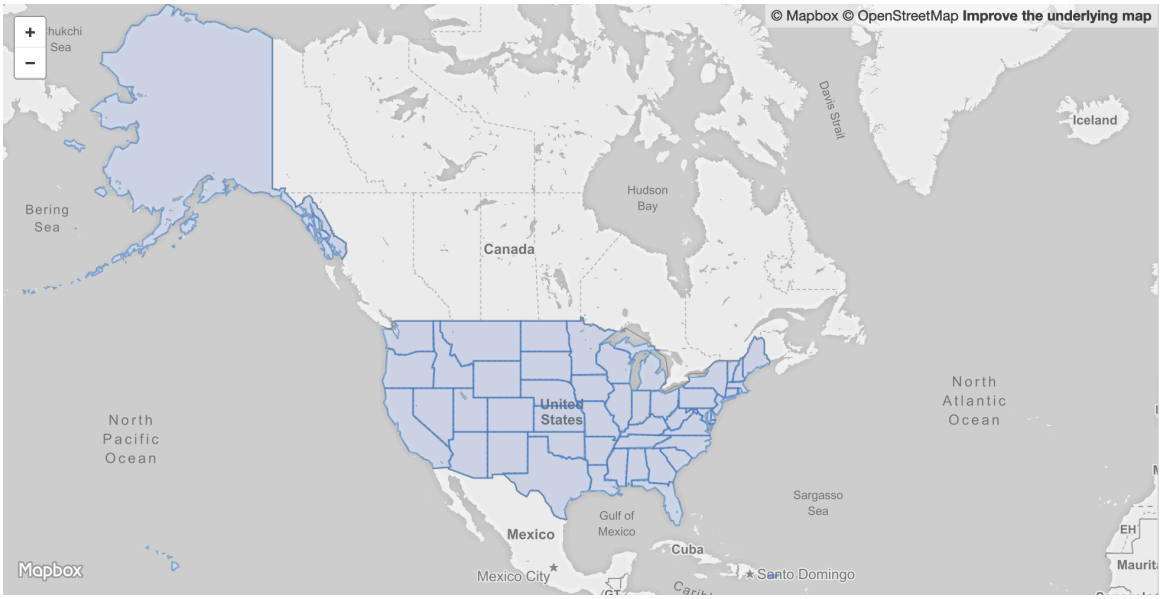


FIGURE 1: SHOWING THE RENDERED STATES WITH BOUNDARIES ON A WEB-BROWSER

Part -I Exploratory Data Analysis

A BRIEF INTRODUCTION TO EDA

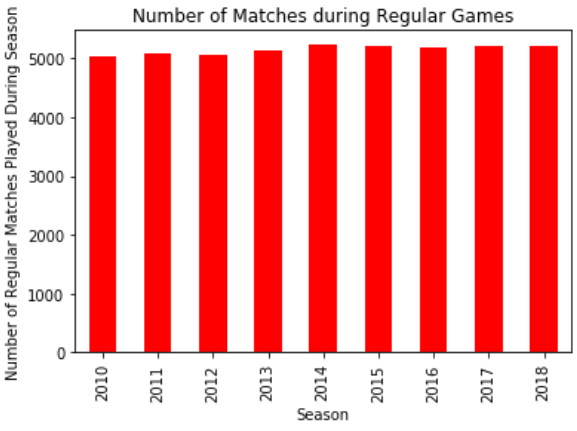
The first step in any investigation is to develop an understanding of what the data truly represents, and in the process, attempt to unravel the hidden "stories" in the data. This step is crucial as attempting to directly ask questions without completely understanding the dataset in is almost always hazardous! There are many techniques used today to explore datasets. However, the one that we used while exploring the NCAA Women's Basketball Games Dataset is that of *Exploratory Data Analysis*, more commonly referred to as EDA today. In the words of John Tukey, (*who literally wrote the book on EDA in 1977)*, The primary objectives of EDA are multifold and are as follows:

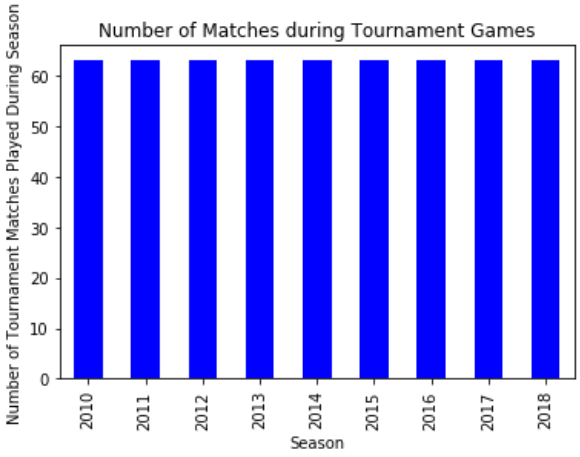
1. To Suggest Hypothesis about the causes of observed phenomena.
2. Assess the Assumptions on which Statistical Inference will be based.
3. Support the selection of appropriate statistical tools and techniques.
4. Provide a basis for further data collection through surveys or experiments.

With this in mind, we will now explore the dataset/s which are provided for us from the Kaggle website. We will not go into in-depth statistical analysis, but rather explore the kinds of questions we can analyse from the data provided for us.

NUMBER OF MATCHES BY-SEASON

The first question we asked ourselves was if the number of matches played during each season (from 2010 through 2018) is constant. From the underlying structure of the dataset as described in the Data Description Section, we know that the number of matches played during each season must be roughly constant. However, plotting a bar-graph with the number of matches played during each season will help us to validate the claims. This is because the Season can be traced a s categorical variable! On observation, we notice that the tournament data is in a separate dataset. Thus, we generate two bar-charts with the total number of matches played during each season for regular season and tournament seasons.



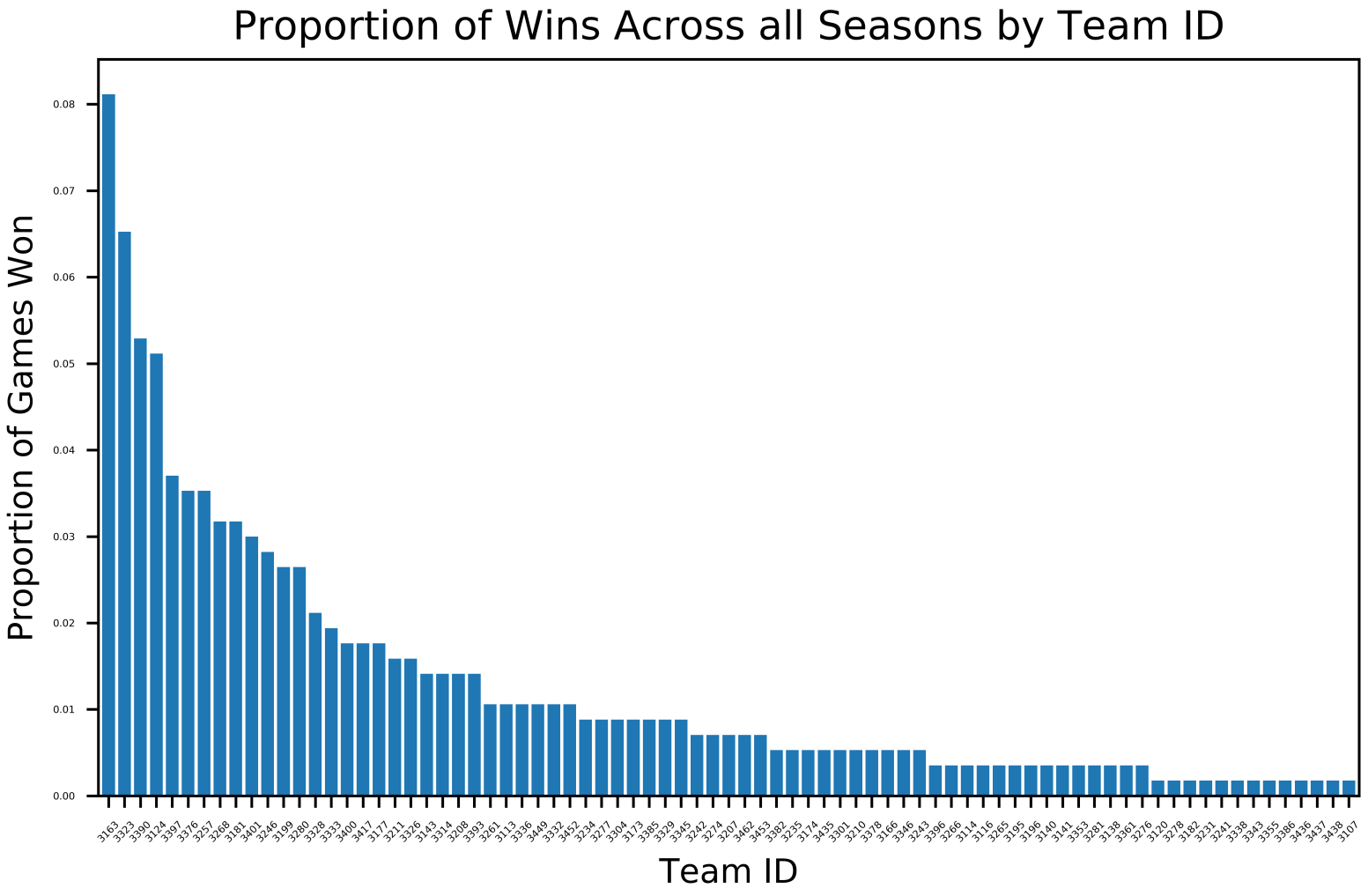


NUMBER OF MATCHES PLAYED ACROSS SEASONS

We can observe quite clearly from the above bar-charts that the number of matches played during the regular season across seasons is nearly constant, and that the number of matches played during the tournament across seasons is a constant number of 63. Thus, the bar charts above validate our claims!

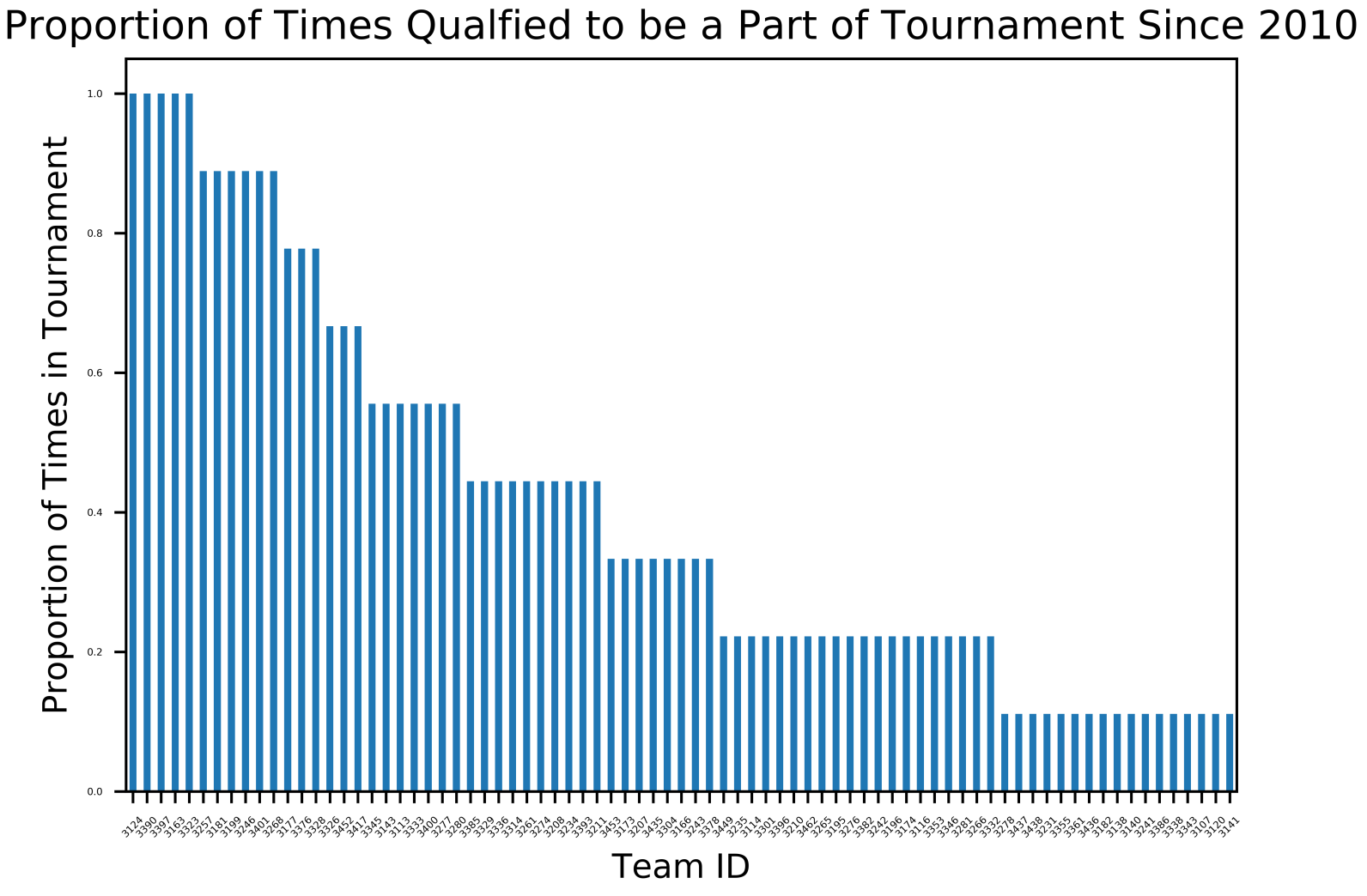
THE BEST TEAM

Another question we asked ourselves was "What is the best team in the NCAA Women's Basketball League?" Answering this question is slightly more complicated. We decided to plot the average proportion of wins in tournament games across seasons, since the top 64 teams from each season are only chosen to play in the tournament. Thus, winning games in the tournament should be harder since only the best compete with the best! To do this, we first retrieve the team name from the *WTeams* file and execute a table-join, in order to correctly map each Team ID to the actual name of the team. Once we join both the datasets, we can now group by each unique Team and then calculate the proportion of the matches won by each team across seasons in tournament games. Since team-names are categorical variables, we use a bar-chart to visualise these results as well. Although we join to obtain the team name for each team code, we use the team code as the x-axis labels, since the variability in the size of names of the institutions is huge, and makes formatting the image harder.



From the above image, we can see that the performance of teams in the tournament season is not uniform. In the bar-plot above we can see that the Team ID with ID 3163 has a considerably higher win-rate across the seasons. However, the Team ID with ID 3107 has won a very small proportion of the games. Cross referencing with the *WTeams* dataset, we find that the team corresponding to 3163 is Connecticut, and the team corresponding to 3017 is Albany NY.

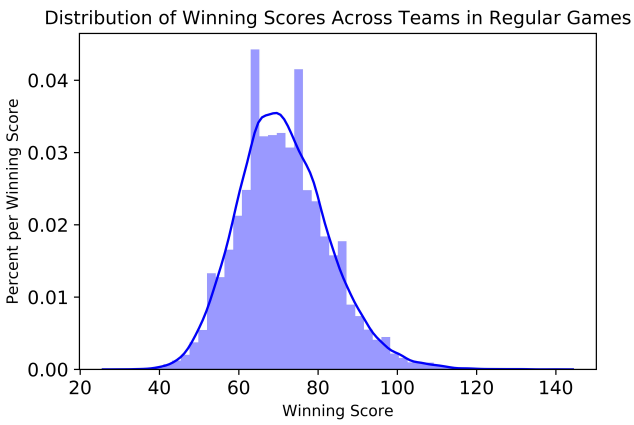
However, does this mean that Connecticut is the best team, and correspondingly Albany NY the worst? The short answer is No! What if Albany NY had played their first season in 2018, and did not have as much experience? Conversely what if Connecticut had played for only the 2018 season and performed extremely well? *(In reality Connecticut is by-far the best team amongst NCAA Women's Basketball teams and Albany is in-fact lower-ranked)* However, we do need to validate these claims in our dataset. Hence, we decide to plot the proportion of times the teams have been selected to be a part of the tournament since 2010. This would indicate that the teams who qualifies more number of times have consistently been in the top 64 teams spanning 9 seasons which innately tells us that they are more likely to be better. Thus, we group by each unique season of the tournament and count the number of times each team was present in the list of teams. This yields the following bar-chart, which represents the proportion of times each team was selected in the top 64 teams in the 9 years spanning this dataset.

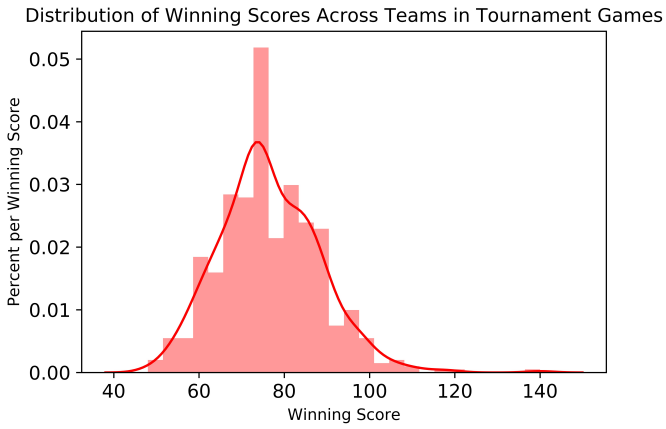


Thus, we can see quite clearly that a sub-section of 5 teams were selected 100% of the time, meaning that they were consistently at the top 64, whereas 17 teams only qualifies once. On Cross referencing with the *WTeams* dataset, we find that the teams who qualified all of the time are Baylor, Connecticut, Notre Dame, Stanford and Tennessee. Thus, we have shown that these teams have a high chance of being better teams in general. However, hypothesis testing and statistical validation will only reveal whether these hypothesis are statistically plausible or not. Further, on comparing the two plots, we can see that Connecticut qualified 100% of the time to tournaments and won 80% of the games on average across the tournaments, suggesting that they are in fact an extremely strong team. Similarly we can see that Albany NY have qualifies only once out of the 9 seasons, coupled with an extremely low proportion of wins. This suggests that they are a comparatively weaker team. But, to re-iterate all of these hypotheses can be concretised with the help of rigorous testing.

DISTRIBUTIONS OF SCORES

Now, we have an understanding of which teams consistently perform in the series, it is important to investigate is the distribution of final scores from each games differ from the regular games to the tournament games. From a logical standpoint, we would expect that there would be a greater viability in the scores during the regular season, since it could be possible that a very strong team is paired with a substantially weaker team such as *Connecticut* and *Albany NY*. In order to analyse these differences the first step is to plot histograms, with embedded kernel density estimates for the distribution of scores over the regular season and the tournament. The embedded KDE is in bold, with a translucent histogram for the regular seasons and tournament seasons. It must be kept in mind that the histograms describe the overall distribution of winning scores over the 9 years between the 2010 season through the 2018 season. The histograms generated are below:

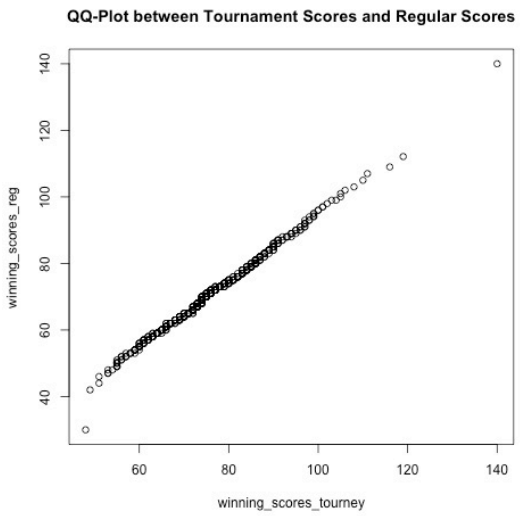




From the blue histogram, we can see that the distribution of scores for the regular games from 2010 though 2018, is a bimodal distribution. Further, the mean of the distribution is 71.28 points. The first peak corresponds to somewhere around 60, and the second peak corresponds to somewhere around 75. Further, the distribution has little to no skew as well. This is validated when the skew is computed to be roughly 0.448, thus suggesting a near symmetric distribution. The kurtosis of the distribution of these scores also happens to be approximately 3.42, which suggests that the data is roughly normal. This is also validating

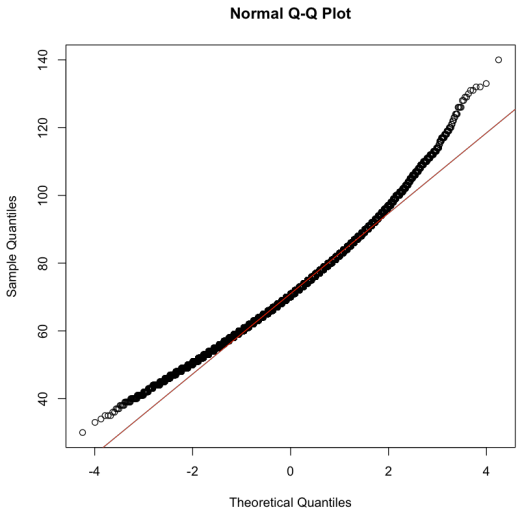
the visual KDE which shows a distribution which is roughly symmetric with little/no skew estimated through non parametric techniques.

Similarly, from the red histogram, we can see that the distribution of scores for the tournament games from 2010 though 2018, is a unimodal distribution. Further, the mean of the distribution is 76.53 points. Further, the distribution has a slight right skew as well. This is validated when the skew is computed to be roughly 0.5725, thus suggesting a slightly right skewed distribution. This is also again validated by the visual Kernel Density Estimate which is plotted on the histogram in a bold red line. The KDE shows a deviation at around a winning score of 85. The kurtosis of the distribution of these scores also happens to be approximately 4.39565, which suggests that the data deviates from a normal distribution. This is also validating the visual KDE which shows a distribution which is has a slight right skew estimated through non-parametric techniques. From this analysis of the histograms of the distributions of the two scores across the regular games and tournaments, we can see that they are indeed different. However let us plot a *Quantile Quantile* Plot between the two distribution to see whether they are indeed different.

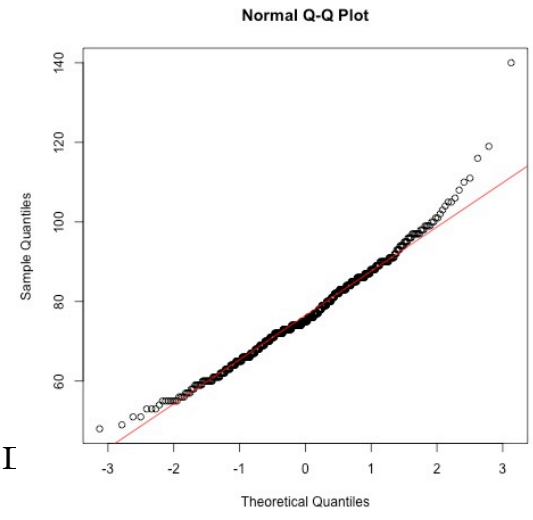


The histograms obtained are shown on the left. The first histogram is a QQ-plot which is drawn between the tournament scores and the regular seasons scores. From the QQ-Plot below, we can see that the quantile-quantile plot between the regular season and tournament scores are roughly linear. In fact the relationship is strongly linear and this suggests that the distribution of scores from the tournament and the regular season were in-fact similar. This result goes against our intuitive notion and the obtained histograms above and is rather surprising.

Following this unusual result, we then plot the quantile quantile plots between the normal distribution and the tournament and regular season scores to see whether they agree with the conclusions from our histogram plots.



E.I



The QQ-Plot on the left shows the QQ-Plot of the quantiles for the regular scores for the seasons with the corresponding quantiles for a normal distribution.The QQ-Plot for regular season scores shows deviations from the normal distribution at extremely high values of the theoretical quantiles, Further, the relationship between the theoretical quintiles and actual quintiles is in fact *non-linear* as suggested by the embedded straight line drawn between the first and third quantiles. The QQ-Plot for tournament scores on the right shows much greater deviations from the normal distribution at extremely high values of the theoretical quantiles, however once again the relationship between the theoretical quantiles and actual quantiles is in fact a*pproximately linear* especially in the quantiles ranging from -2 to 2.

The conclusions from the QQ-Plots suggest that the tournament and the regular season scores can are in fact similarly distributed, however, while the tournament scores are more likely to be normal, the regular season scores shows clear deviations from normality.

Since this conclusion goes against what we developed in the histograms, we run a Kolmogorov-Smirnov 2 Sample test under the null hypothesis that the 2 datasets os scores were generated from the same distribution. We set a significance level of 0.05 for this test. In other words, we reject the null hypothesis if the p-value is less than 0.05. The test suggests that the p-value is roughly 1.196697625790175e-23 with a D statistic of 0.21733032769097965. This suggests that we can reject the null hypothesis, that the 2 distributions came from the same distribution.

Thus, the ultimate conclusion is that the data suggests that the distribution of scores during the tournament and the the regular season is in fact different. However further investigation regarding the discrepancy between the conclusions offered by the QQ-Plot and the K.S 2 Sample test and the histograms must be undertaken potentially by collecting more data. The reason as to why the distribution of scores may be different must be studied with the help of an in-depth analysis as well.

LOCATION OF WINS

Now that we have realised that the distribution of scores between the tournament and the regular season could potentially be different, let us investigate whether the location of a game influences the win rate across teams. To do this we use the Regular Games and Tourney Games Datasets. We observe that there is a field labelled WLoc which takes on one of 3 values. ‘H’ for Home, ‘A’, for Away *(The home location of the opponent)* and “N” which stands for a neutral location.

Upon aggregating the total proportion of wins across each category of location, we observe that the proportion of wins at home at the tournament and the regular season is substantially higher than away as well as neutral locations. This consistent observation across the tournament as well as the regular season may be an indication of the well-known “*Home Court Advantage”*. However, a tabular summary alone would not suffice to conclude as to whether a true home court advantage exists and extensive statistical testing in a team by team level and a tournament level is necessary in order to validate any claims.

location	Win
A	0.388655
H	0.611345
N	0.500000

Table I. Regular Season.

location	Win
A	0.152284
H	0.847716
N	0.500000

Table II. Tournament.

Another key factor consider while consider is the actual "location" of the wins themselves. In our case the *WGameCities* and the *WCities* Dataset provide information on the actual location of the matches. However extensive modifications were made to the dataset which are highlighted in the box below:

PRE-PROCESSING THE DATA

1. The *WGameCities* contains details on all games since the 2010 season irrespective of the kind of tournament (that is it combines both the tournament and the regular seasons). The first step is, to map the City and State to the City ID as provided in the *WGameCities* Dataset.

2. To do this, first, we concatenate the City and State in the WCities.csv file with an '@' to avoid wrongly mapping an incorrect City with a Game in the WGameCities file.

3. We can then merge the combined City@State Column with the CityID column to obtain the location of the various games in the WGameCities dataset. The, we split the City@State into the City and State columns in order to facilitate further processing.

4. We then merge the WTeamID and the LTeamID from the Tourney and Regular Detailed Datasets in order to obtain the actual name of a team rather than a team ID.

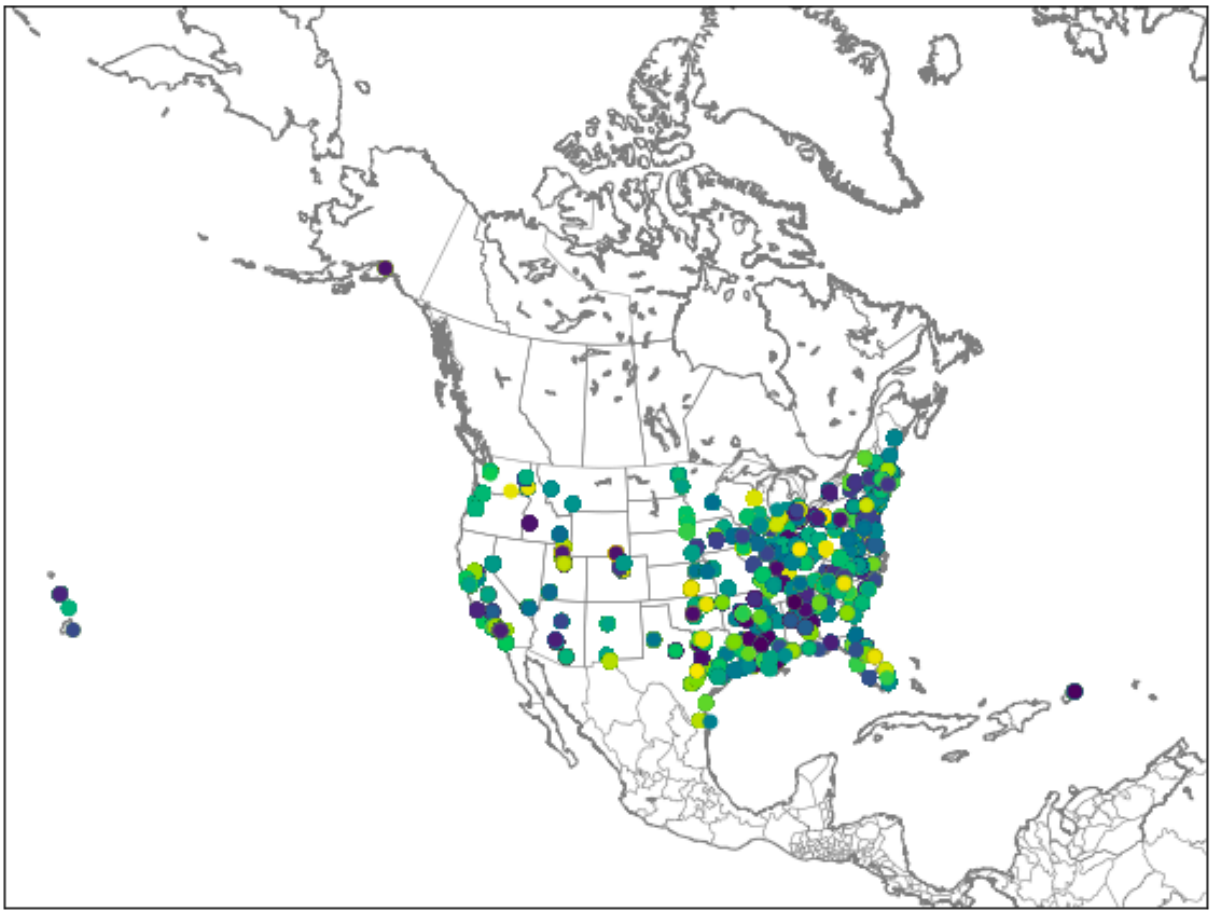
	Season	DayNum	WTeamID	LTeamID	CRType	City	State
0	2010	11	Akron	IUPUI	Regular	akron	OH
1	2010	30	Akron	IPFW	Regular	akron	OH
2	2010	33	Akron	Youngstown St	Regular	akron	OH
3	2010	56	Temple	Akron	Regular	akron	OH
4	2010	68	Akron	Buffalo	Regular	akron	OH

5. At this stage the WGameCities Dataset will appear like above. Finally, now we use an external dataset which is publicly available to map each city state pair with a let of latitude longitude pairs. This would allow us to geographically map the distribution of wins.

	Season	DayNum	WTeamID	LTeamID	CRType	City	State	lat	lng
0	2010	11	Akron	IUPUI	Regular	akron	OH	41.0802	-81.5219
1	2010	30	Akron	IPFW	Regular	akron	OH	41.0802	-81.5219
2	2010	33	Akron	Youngstown St	Regular	akron	OH	41.0802	-81.5219
3	2010	56	Temple	Akron	Regular	akron	OH	41.0802	-81.5219
4	2010	68	Akron	Buffalo	Regular	akron	OH	41.0802	-81.5219

6. At this point, we are ready to finally map the distribution of wins for various teams across various locations for both the regular and tournament datasets since the WGameCities Dataset contains all information within a single dataset.

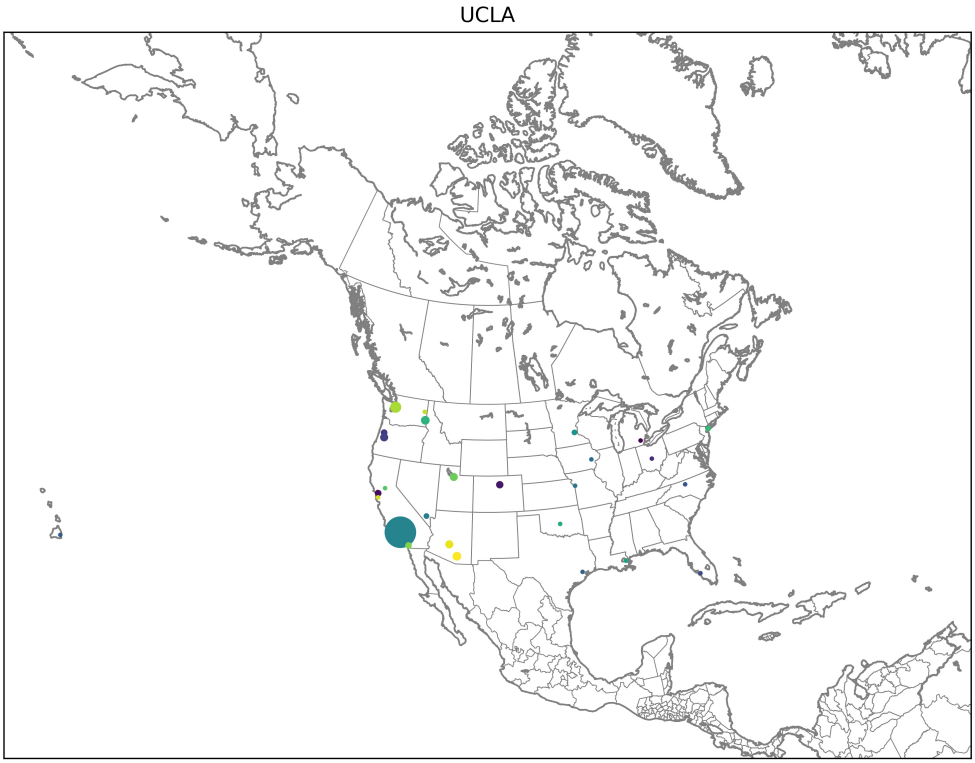
In order to facilitate Geographic Mapping, we use *BaseMap*, which is an addition to the native Matplotlib library in Python. The library facilities the usage of cartographic projections to visualise the spherical surface of the Earth on a 2 dimensional screen as well. We approximate the location of the game to the latitude and longitude in which a city is located. This is because getting the actual coordinates of the arena in which a game took place is a cumbersome process involving scraping and cross referencing between multiple datasets, which can ultimately increase the likelihood of making errors while cross-referencing between the various datasets. The below map shows the geographic distribution of matches played all over the U.S across tournaments and regular seasons from the 2010 to 2018 season.



GAMES PLAYED ACROSS THE U.S FROM THE 2010 SEASON TO 2018 SEASON

We can see that the matches played are not evenly distributed across the U.S. There seems to be a greater concentration of games played in the Mid-Western, Eastern, and Southern United States as opposed to the Western and Northern United States. This may suggest suggest that basketball games are biased to play in those locations. However, as always rigorous statistical testing is necessary to test/validate any of these claims!

Using this, we can also generate separate maps for each institution to visualise where they play most of their matches! Further, in order to test our claims of a potential home court advantage, we have scaled the size of each scatter plot by the proportion of games won by the team at that particular location. As an example, we have visualised the distribution of wins for UCLA across various locations, scaled by the proportion of games UCLA won at that location, from the 2010 to 2018 season below! however feel free to check all maps for every school here !



WINS FOR UCLA ACROSS THE U.S

E.D.A

We can see quite clearly that the distribution of wins for UCLA is not uniformly scattered. IN fact, we can see a huge increase in a location in Southern California (Most Likely L.A). Further, we see that there is a decrease in win rates as we move from West to East (in other words Away From Home). Finally, we see that the win rates at locations closer to L.A, is higher than those away from them. The map above thus, provides two fundamental questions for us. First, Is there a Home Court Advantage after all? Again, We need to validate these claims only with additional analysis, however, the tabular analysis earlier and the geographic analysis above clearly provide good motivation to analyse these through statistical validation techniques. Second, given the match happened at a Neutral Location, is the location really "neutral". That is are there any external influences which may bias the neutrality of a location toward a specific team. For instance, if a match occurred between UCLA and Kansas State at San Diego, is UCLA more likely to win the match due to its close proximity to San Diego? The geographic analysis above showing the decreasing sizes of the scatterplots as we move away from the home location good motivation to analyse the same through statistical validation techniques.

GAMES BY REGION

In the first geographic visualisation, we saw that the location of matches were not uniformly distributed all across the United States. Further, in the second graph we saw that the distribution of wins for each teams is not uniformly distributed across the U.S, and in fact has higher rates closer to the potential home of the team. This poises a natural question that, are the proportion of matches played equal across various geographic regions in the U.S? That is, for instance is the proportion of game played in the Mid-western United States the same as teams in the Western Region? Well, in Order to answer this question, we first realise that the details regarding the geographic region is not inferable from the dataset owing to its structure. In other words since each region was randomised during each Season, we need to make use of Official U.S Census Bureau Datasets in order to classify the location of the home location of the particular team.

FINDING THE HOME LOCATION OF EACH TEAM

1. First we need to get the actual home location of each team. To do this we cross-reference between the *WRegularSeasonDetailedResults* Dataset containing the WLoc field which states either 'H', 'A' or 'N'. Further, since the structure of the data allows us uniquely identify each match with the help of a unique combination of Winning team, Day Number and the Season, we can then map the WLoc Field to the *WGameCitiesDataset*, and thus obtain a field WLoc which says 'H', 'A' or 'N' in the WGameCities Dataset.

2. Finally we can drop duplicate teams and thus have only unique teams, and map each 'H' WLoc to the Location in which the Match was played and thereby in-turn map it to the team who won the team. This ultimately can create a Python dictionary containing the Team and Home Location for <u>most</u> of the teams in the dataset.

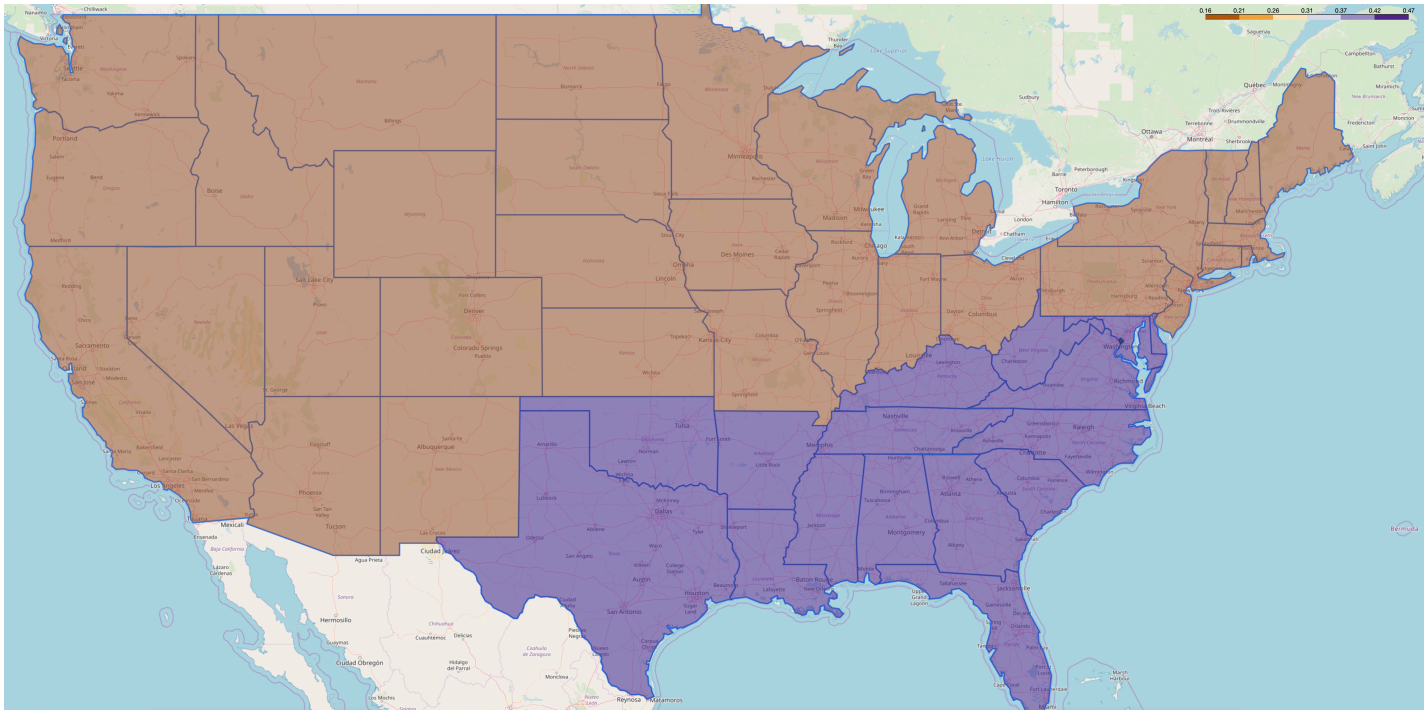
3. After finding the home location, we can now use the U.S Census Bureau Dataset to find the official region in which a team is based, by using the home location of each team.

Finally, we can now group-by the region, and calculate the proportion of wins across all teams for each region within the U.S. The resulting dataset will be something like below!

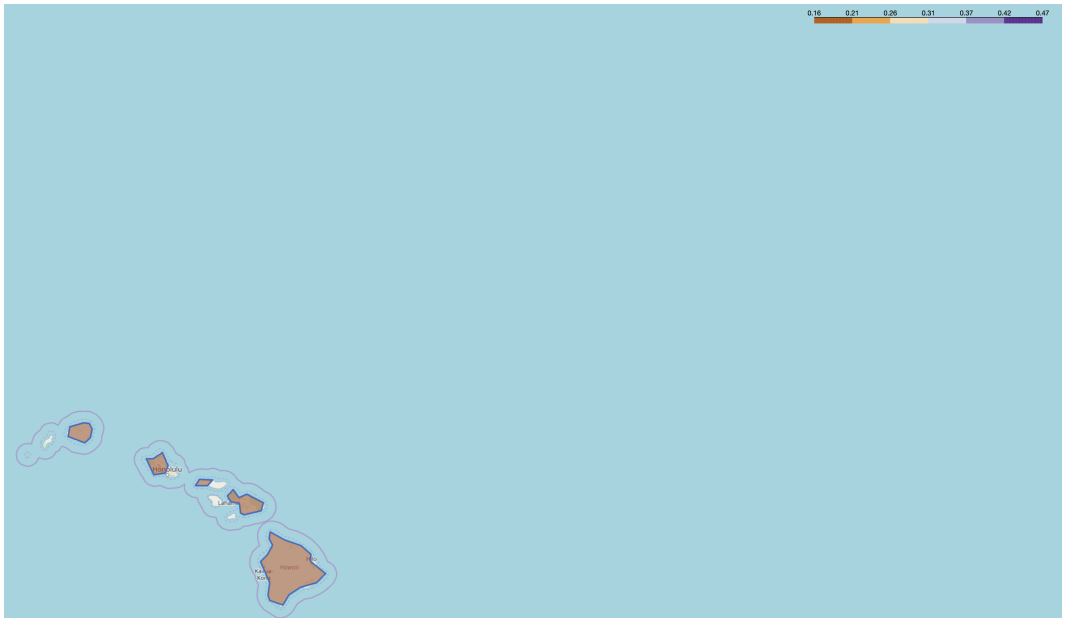
	Season	DayNum	WTeamID	LTeamID	CRType	City	State	lat	lng	Region	prop_matches
0	2010	11	Akron	IUPUI	Regular	akron	OH	41.0802	-81.5219	Midwest	0.190801
1	2010	30	Akron	IPFW	Regular	akron	OH	41.0802	-81.5219	Midwest	0.190801
2	2010	33	Akron	Youngstown St	Regular	akron	OH	41.0802	-81.5219	Midwest	0.190801
3	2010	56	Temple	Akron	Regular	akron	OH	41.0802	-81.5219	Midwest	0.190801
4	2010	68	Akron	Buffalo	Regular	akron	OH	41.0802	-81.5219	Midwest	0.190801

Once we have found the win rates for each region, we then group by each region and count the number of matches played in that region from one season to another. Finally, we use a choropeth map which has the proportion of matches played during the Tournament and the Regular season during the 2010 to 2018 NCAA Women's basketball season. The choropeth map is aded via a geojson file which contains the various state boundaries across the U.S. Finally we also colour code the choropeth map by the proportion of game played in that region. The visualised results are generated using *Folium*, which is powered by Leaflet, and is a .html file. The rendered results in a web-browser are shown below.

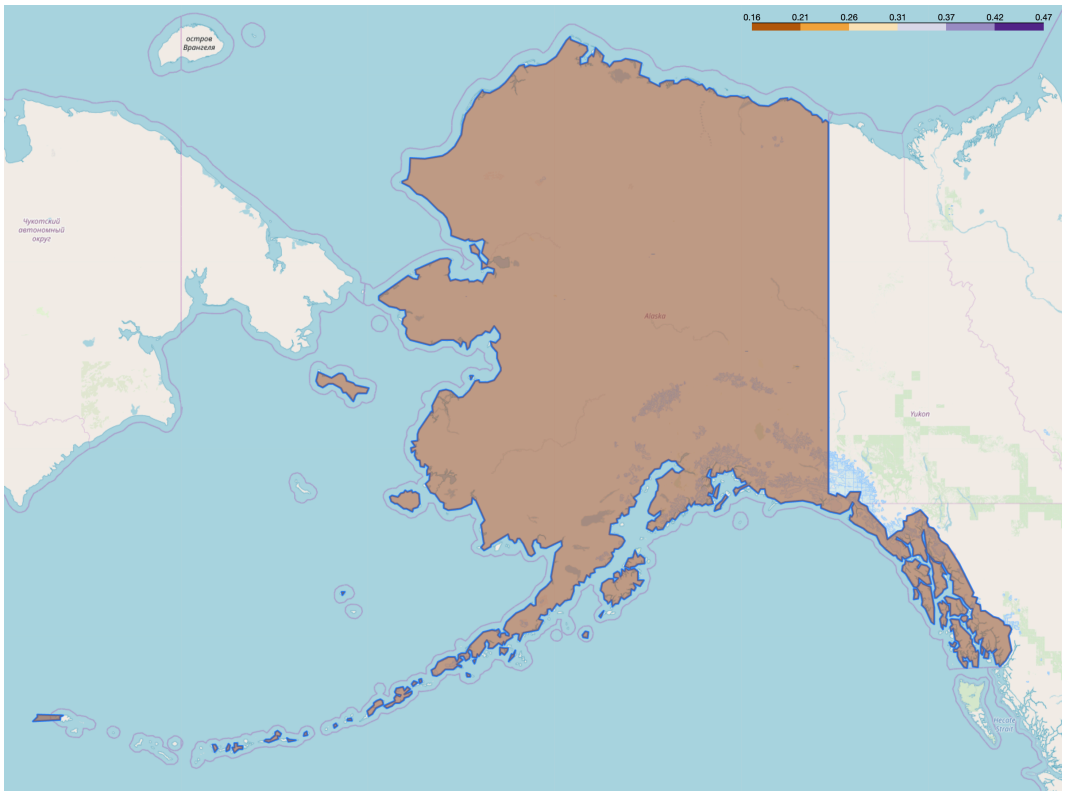
PROPORTION OF GAMES PLAYED BY REGION ACROSS THE CONTINENTAL U.S FROM 2010 THROUGH 2018.



PROPORTION OF GAMES PLAYED BY REGION ACROSS HAWAII FROM 2010 THROUGH 2018.



PROPORTION OF GAMES PLAYED BY REGION ACROSS ALASKA FROM 2010 THROUGH 2018.



As we can see quite clearly from the graph, proportion of matches played across the North-Eastern, Mid-Western, and Western United States is approximately constant, and is roughly 15%, whereas the proportion fo games played in the Southern United States is much higher and is approximately 45% as indicated by the purple colour on the graph. Further, the proportion of games played in Alaska and Hawaii is also shown separately since the total map could not fit in a single screenshot! The proportion of matches across Alaska and Hawaii is the same because of the fact that both, Alaska and Hawaii are classified as a part of the Western Region by the U.S Census Bureau. The reasons for the unequal distributions can be attributed to potentially, the higher popularity of basketball amongst people in the Southern United States. However, such suggestions are merely speculations and must be validated with the help of statistical testing to have any meaningful conclusions.

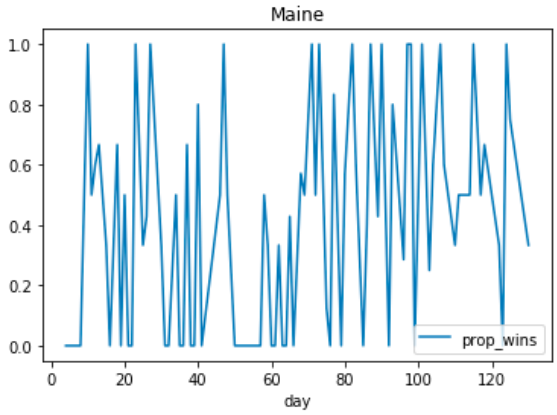
WINS OVER TIME

While the list of questions we can ask ourselves is endless, let us ask one final question! How do the wins of each team vary with the day number of the series. Yes, the days are not "same" for each season, since we have observed that the start date of each season (the day of the first match) does vary. For instance, during some seasons the first game is played only on day 4 of the season! Further, for the women's Basketball data in particular we see that the is even more variation in the days of the tournament. This hinders parallel comparison of win dates by dates. However, since the extent of parallel comparison is higher for the regular dataset, and since it involves a greater number of teams, we choose to perform analysis on the regular dataset. However, the extent of pre-processing required is extensive for this analysis.

To do this, we first group by each day of the season. Then we extract the Pandas series containing the winning teams during that day of the series, and take the proportion of the values which were that particular value. This would contain the proportion of wins for that particular team on that particular day. Then finally, using Python Dictionaries we aggregate the win proportions for each team indexes by each day of the series.

Finally, to visualise our data we then plot a line plot which is ordered chronologically on the x-axis. That is, it begins with day zero. Upon plotting such a graph for each institution, we see the following:

Consider the Line-plot for Maine:

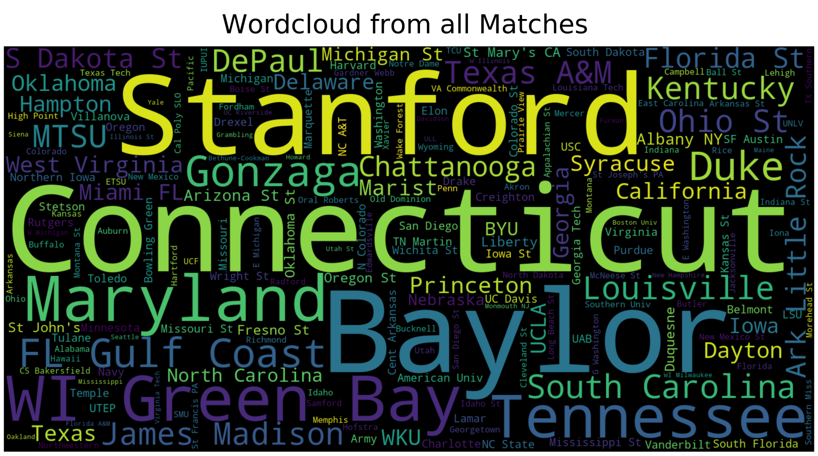


From the line plot toward the right we can see that the performance for Maine (measured by the proportion of games won that that day over the 9 year span over which the data was collected) seems to have to relationships. In other words it appears completely random. Further, the implicit problems in not having truly parallel timelines from one season to another makes any comparison between teams very hard to make. For instance, we can never answer if Maine or any other school for that matter, performs better as the

number of days in the season progresses, or in other words answer the question, “How does a particular school perform under pressure” since, as the days progresses there is an increasing pressure to do well, in order to emerge within the first 64 teams. However the non-parallelism between the timelines of regular season from one season to another makes any meaningful comparison impossible. This effect is even greater for the tournament dataset, since there are greater variation in the start dates and analysing these time-series based trends must be done with most complicated and sophisticated analysis techniques.

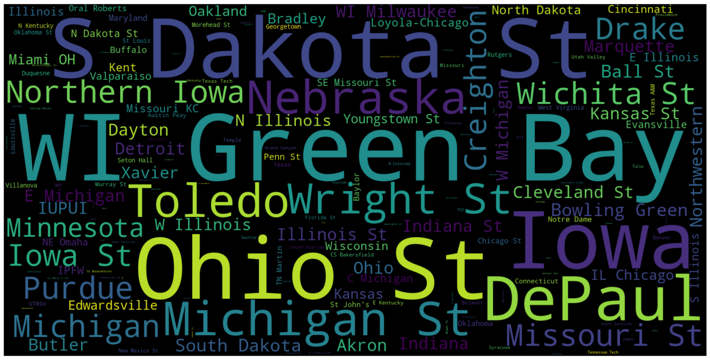
REVISITING THE BEST TEAM

We saw before that the best team within the NCAA Women’s basketball data was hard to quantify. Further, we had an inconvenient approach of plotting the histograms by the ID of the team, and had to neglect the regular game data while performing our analysis. However, we can look at the best team from another Angle by considering Word Clouds. Since we have now obtained the name of each team, we can now use word clouds, and the WGameCities dataset to see whether our conclusions from the previous bar-chart based analyses hold true.



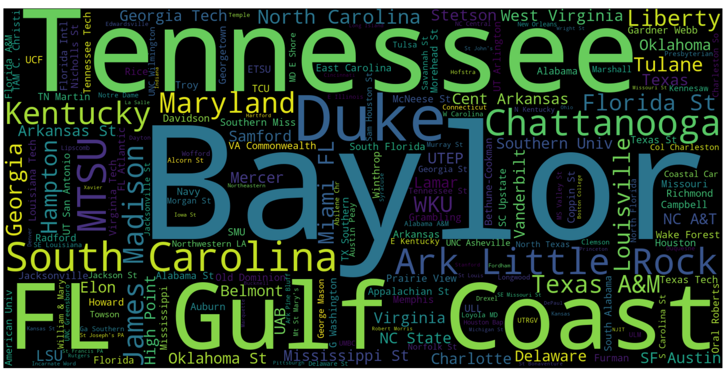
The word clouds help us to visualise our data in a compact manner. The strategy used to scale the size of each team name is the proportion in % of matches won in both Tournament and Regular matches over the 2010 to 2018 season. We observe the following Word Cloud. The following word cloud shows us that the most successful teams overall are Stanford, Connecticut and Baylor. These teams are followed closely by Teams such as Maryland and Green Bay. The results from the above word cloud do in-fact agree wit the results in our histogram analysis, thus suggesting that our claims earlier were not completely unfounded. Finally, as a result of the data pre-processing, we can also generate separate word clouds for each region as classified by the U.S Census bureau. This would show the regional “champions” in the data. This comparison is not incorrect to make as the regions in which the teams are based do not change, however, the naming of each region varies randomly between W, X, Y, or Z in the NCAA Women’s Basketball data. The word clouds by region are given on the next page

Wordcloud from the Midwest



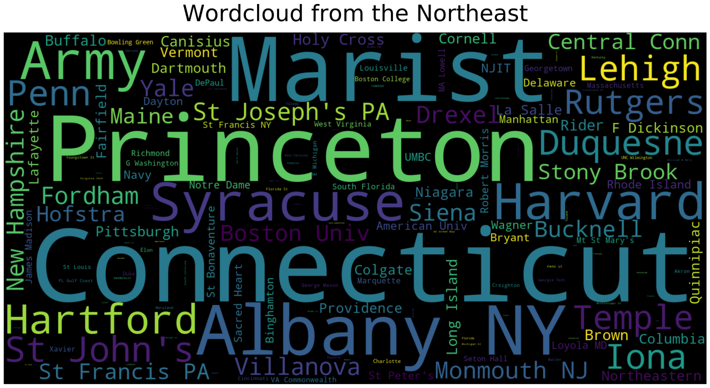
We can see in the Midwest that South Dakota State, WI, Green Bay, Ohio State are the top teams in this region. Further, we see that there are a large number of teams who have slightly smaller, but sizeable number of wins as well. This could suggest that the Midwest is a particularly competitive region to play in owing to the large proportion of teams that perform well.

Wordcloud from the South



We can see in the South that Tennessee, Baylor, Florida Gulf Cost are the top teams in this region. Further, we see that the size of Baylor is particularly larger than the rest of the teams, thus suggesting that Baylor could indeed be a powerful team! Further, Tennessee and FL Gulf Coast are smaller but similarly sized as well. However the variation between the sizes of the clouds are much higher in the Southern region, indicate a presence of few teams that perform really well and dominate the games in the region.

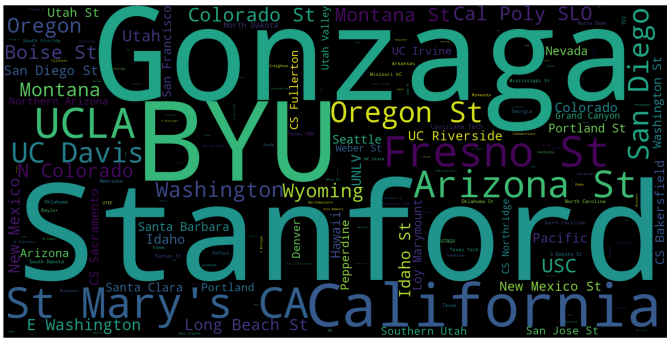
Wordcloud from the Northeast



We can see in the Northeast that Marist, Princeton and Connecticut are the top teams in this region. Further, we see that the size of Connecticut is particularly larger than the rest of the teams, thus suggesting that Connecticut could indeed be a powerful team! Further, Tennessee and FL Gulf Coast are smaller but similarly sized as well. However, we also see that Albany NY, the team from out previous

analysis is also present here! Further, we see that it has a similar size to Harvard in terms of proportion of games won. This could suggest that Albany NY is in fact a regional power, and that it loses out to significantly more powerful teams such as Connecticut, which from our precious analysis seemed to be one of the most successful teams in the NCAA Women's dataset.

Wordcloud from the West



Finally, the word cloud from the West region is in-fact really interesting. The top performing teams from the region happens to be Gonzaga, Brigham Young University and Stanford. It is interesting to see that in the West Region, the scaling of the sizes of these teams has high variation amongst the teams. This could once again suggest the presence of a few highly powerful teams that dominate the region something on the lines of the South Region. However, the variation in word sizes in the West is even greater than the South thus, suggesting an even greater effect in the region. Finally, it is interesting to note that there are no teams from the outlying states (Hawaii and Alaska) represented having high team sizes here at all.

Thus, the word-cloud analysis is extremely invaluable in our analysis as it helped strengthen some of our hypothesis on the identification of the beast team from the NCAA dataset, while at the same time providing a compact and elegant way of summarising a vast amount of data! Finally the observations from this study must be validated with statistical approaches in order to further strengthen these hypotheses.

SUMMARY

Finally, let us conclude the Exploratory Data Analysis Segment of this Investigation. Through this segment, we have, at various stages, suggested possible causes for various observed phenomena in the dataset, assessed if conditions for meaningful statistical conclusions are present in potential questions in the dataset, supported our observations with the help of statistical techniques and visualisation and finally also provided cases where additional data maybe be needed to form stronger, statistically valid conclusions from the data. Thus, are have satisfied Tukey's four pillars for a successful EDA process!

From the EDA we have learned from the data that:

1. The number of matches played from one season to another is in-fact roughly constant.

2. Identification of the "best" team from the NCAA Women's basketball dataset is not as easily as it may seem as there are multiple ways of quantifying how good a particular team is, and that analysis must be done by cross-validating between multiple given datasets, in order to make a potentially statically valid conclusion. Finally, generating Word-clouds later helped us to strengthen some of our hypothesis from the earlier analysis, and frame better hypotheses from the data through an elegant and compact analysis.

3. The distribution of the number of wins varies from the regular series to that of the tournament. This was validated using extensive statical techniques, thus suggesting a very strong conclusion for the same.

4. The distribution of wins varies across locations as well. Through graphical visualisation techniques and geo-spatial visualisation techniques we have seen that the win-rates for teams seems to depend on the location in which the game is being played. The presence of a "home court advantage" is also plausible, however all of these observation must be quantifies through statistical techniques to have any meaningful effect!

5. The proportion of matches played across the U.S also varies! In fact, the Southern region has a much greater proportion played across since the 2010 season. However, these observations need to be validated with the help of rigorous statistical analysis to provide a basis for any further analysis.

6. Comparison of wins across time is not possible in this dataset despite having a seemingly "parallel" timeline because of discrepancies in how the data is finally structured, which finally makes any time-time-series based analysis extremely complicated and challenging!

Now that we have generated a list of interest-areas, we can analyse a few of these questions and try to validate our observations using statistical techniques.

Part -II The Home Court Advantage

The home court advantage is a well-established phenomenon witnessed across all types of sports. This phenomenon occurs when a team performs better overall at its home court than at its opponents' territory. Not surprisingly, the data on the past NCAA seasons yield similar results. Figure I below displays the first ten teams (based on alphabetical order) with their respective win rates for away, home, and neutral games. It can be seen that teams tend to perform better at home.

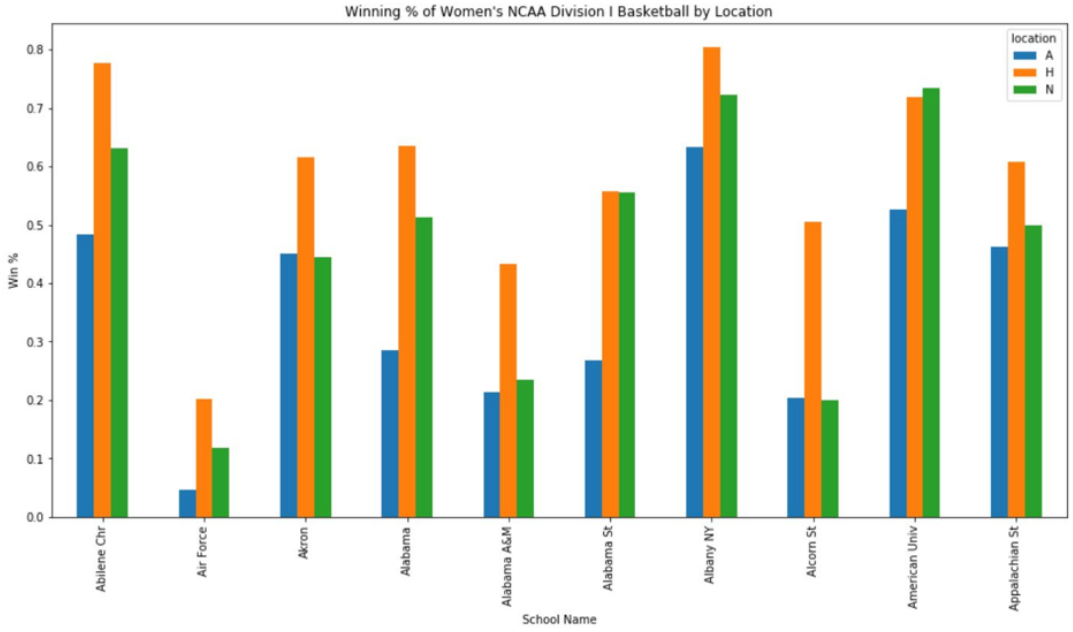


Figure I. Winning rates of teams based on the location of games.

Upon further investigation, it was observed that the disparities between away and home winning rates were quite large. Table I shows the proportions of games won by all teams combined during the regular season, and it can be observed how the winning rate at home is almost twice as large as the winning rate when away. As for the tournament games, the proportion of games won at home is almost six times more than the games won in foreign territories (Table II).

location	Win
A	0.388655
H	0.611345
N	0.500000

Table I. Regular Season.

location	Win
A	0.152284
H	0.847716
N	0.500000

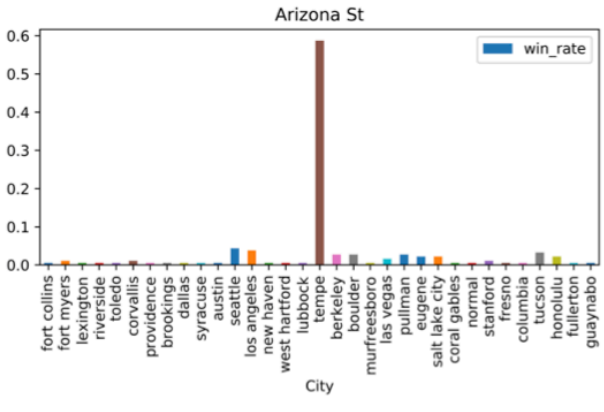
Table II. Tournament.

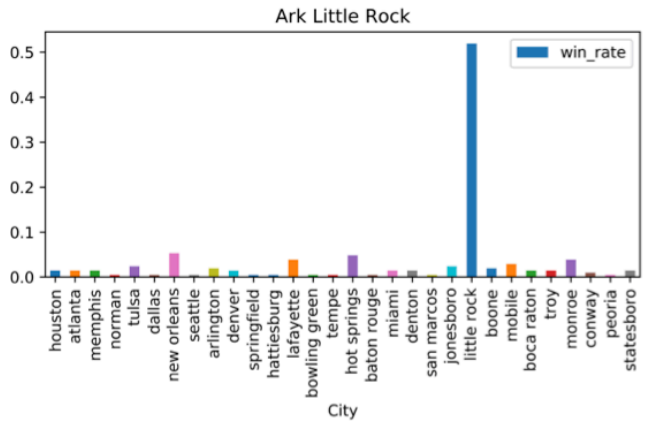
To check whether or not the differences are significant, an idealised method would involve the two-proportion t-test, performed for both regular season and tournament games. This is because the population standard deviation is unknown. However, the degrees of freedom in our case is so large, that the students t distribution approximately converges to that of a normal distribution. Since the sample size was large enough so that the degrees of freedom for performing a t-test exceeds 100, the p-value was approximated with the Normal distribution. In such case, the p-values are $P(z>=64.535) \approx 0$ and $P(z>=13.804) \approx 0$ for both the regular season and tournament respectively. The test results show that the null hypotheses should be rejected. In other words, it is very unlikely that the proportion of home wins being much larger than the proportion of away wins is due to chance. However, this trend only shows a "big-picture". That is, it shows that there is an overall trend to perform better at locations which are their home.

However, what about a school-by-school analysis on the performance over the regular and tournament season games from 2010 through 2018? Hence, by using the *WGameCities* dataset, and the *WNCAARegular* and *WNCAATournament* Datasets, we cross referenced each match by a unique Season, Winning Team ID and City ID. Then we plotted the distribution of wins across each location for each of the teams across tournament and regular season games, having now discovered the home location for each team. Some of the bar-plots we obtained are shown below, we used bar-plots because the wins were distributed over categories. The observations fall into broadly three main categories.

<u>Category-I</u>

The first category (kind) of observations is where there is a single location where there is a large proportion of wins, followed by very few wins at other locations. For instance two such schools were such trends were observed were Arizona St. and the University of Arkansas at Little-Rock.

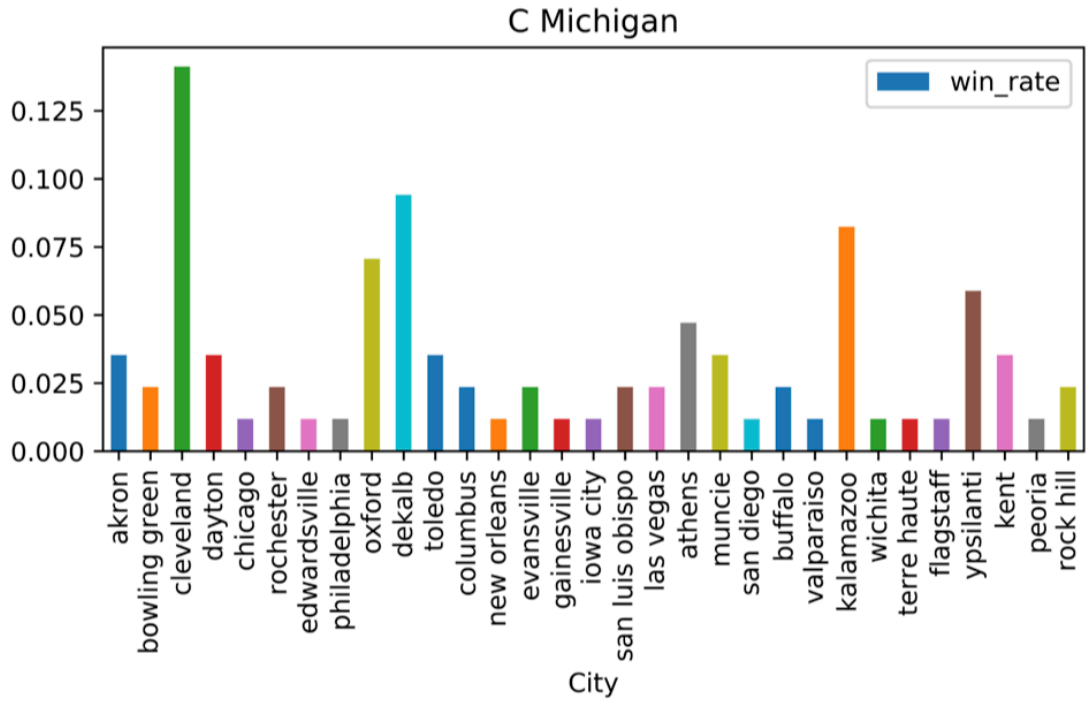




We can see quite clearly from the bar-plots on the previous page that the proportion of wins at Tempe happens to be the home location of Arizona State and Little Rock happens to be the home location of Little Rock (which we have verified programmatically in the dataset, and also know through domain expertise). These schools can be said to have a high degree of home court advantage since a majority of their wins are concentrated at home.

Category-II

The second category of schools are the ones where the location of most wins is not the home location! While rare in this dataset, there are some schools where the maximum proportions of wins for the team in not the home location. One such example of such a school is Central Michigan.

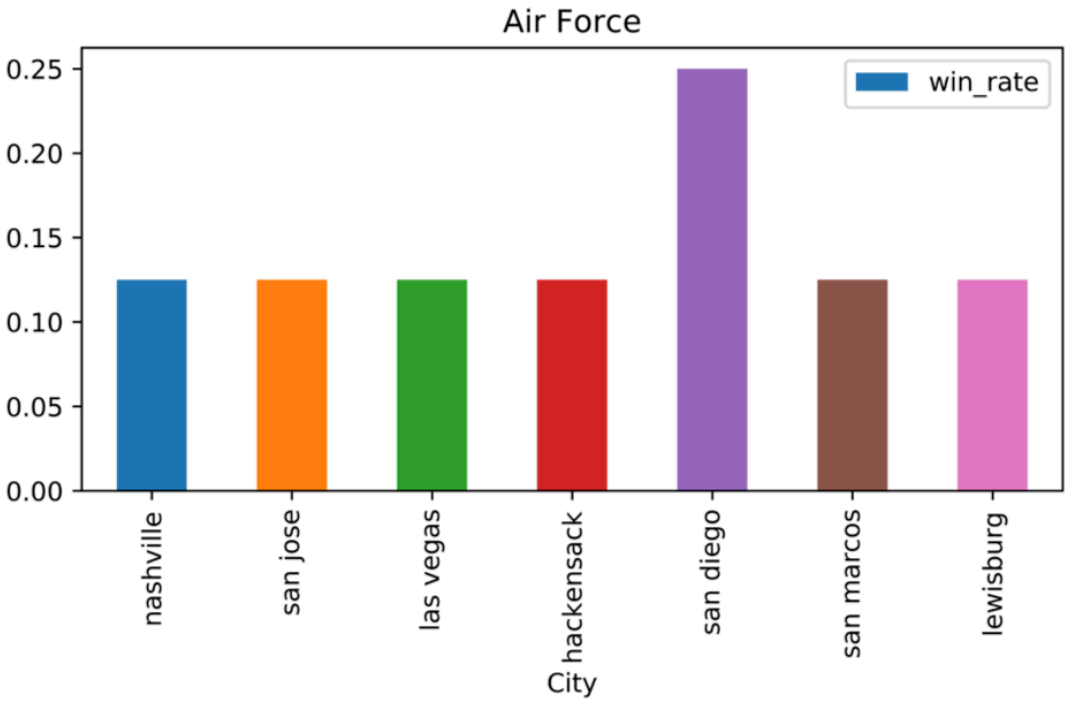


On observation from the bar graph of win proportions, we can see that the proportion of wins is maximum at Cleveland Ohio! Further, we can see that the win rates at other cities are moderately proportioned as well, and there is a more equitable distribution of win proportions across various

locations, especially when compared to that of category-I. These kind of schools visually exhibit absolutely no kind of home-court advantage at all!

Category-III

The third category of schools are the ones where the proportion of wins are nearly uniformly distributed. In other words, for these schools the heights of the bars are nearly equal. A great example of such an example is Air Force.



While the bar-graph of win proportions do not exactly have equal heights, the win rates are almost uniformly distributed over the categories Further, the extent of the home court advantage for Airforce is present albeit at a slight magnitude, however, if there win proportions were in-fact uniformly distributed over the categories, there would be no home court advantage.

Finally from our initial understanding we have understood from visual and theoretical statistical tools that while existence of a home court advantage does seem to be statistically backed to exist on the overall dataset as a whole, an analysis on a school-by-school level seems to suggest

otherwise. In reality, there happens to be a variability in the existence of a home court advantage as well as the ,and the magnitude of home court advantage seems to vary across schools as well.

So now, we try to analyse the presence of a home court analyse in each school by looping through the entire dataset in Python. When the test was conducted on the entire dataset, on a team-by-team basis, 47 teams had insignificant differences during the regular season and 21 teams during the tournament , with a set significance level (α) of 0.05. In order to get a better perspective on who these 47 teams in the regular season and 21 teams in the tournament were. However, on a quick side note, insignificant statistical conclusions for 21 teams for the tournament indicates that roughly 33% of the teams in the tournament did not have a statistically significant home court advantage. In other words almost 1 in 3 team did not have a statistically backed home court advantage during the tournament. Hence, as a result of the conflicting results between the general trend shown by the dataset, and that of each school, we conclude that without further investigation into other features/variables, it cannot be concluded whether or not having home court advantage is the main factor in determining the outcome of a game. However, it must be clearly notes that the presence of home court advantage cannot be completely ruled out on account of the compelling visual evidence and general trend show. By the dataset as a whole.

Part -III Are Locations Really "Neutral"?

OBJECTIVE

We have seen from the E.D.A section of this investigation, that there are three possible locations in which a game can be played. These are Home (H), Away (A), or Neutral (N). From our previous findings on Home Court Advantage, we can see there *may* a general trend in the dataset for teams to perform better at home arenas. However, a critical question to ask is "What about Neutral Locations?" Are neutral locations truly "neutral" in the sense that both teams have to play the game on sheer skill alone, and not rely on the support of the crowd? Logically speaking, if two teams were to play at a neutral location closer to the home court of one of the teams, We would say that that team would have a higher proportion of wins at that location. For instance, if UCLA and Princeton were to play in San Diego, California, we would like say that UCLA has a higher chance of winning since San Diego is in close proximity to Los Angeles! Let us now investigate how "neutral" locations are across the U.S in the NCAA Women's Basketball data from the 2010 Season to the 2018 Season across Regular Games and Tournament Games!

ANALYSIS

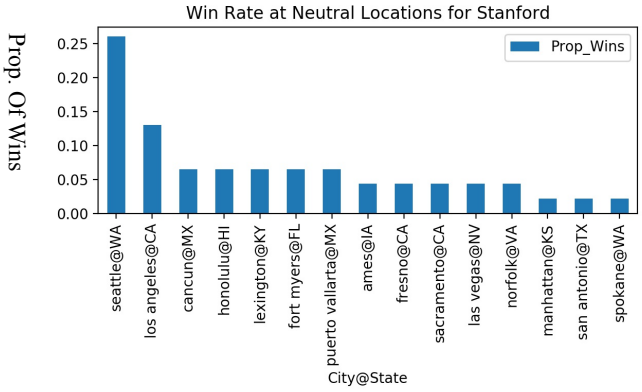
The first step in the analysis is to generate a usable dataset. From which we can examine wins in Neutral Locations. To do this we use the dataset we used during the "Games by Region" EDA Process. In short, we cross-reference between the *WRegularSeasonDetailedResults* Dataset containing the WLoc field which has either 'H', 'A' or 'N'. Further, since the structure of the data allows us uniquely identify each match with the help of a unique combination of Winning team, Day Number and the Season, we can then map the WLoc Field to the *WGameCitiesDataset*, and thus obtain a field WLoc which says 'H', 'A' or 'N' in the WGameCities Dataset. We relabel the newly created 'WLoc' field in the *WGameCitiesDataset* as 'Home'. We then group-by each unique team, and calculate the proportion of wins at each location with label 'N' and create a Python Dictionary keyed by each unique team and valued by a Pandas Series containing the proportion of wins at each "Neutral City". For instance for Akron, here is the value in the dictionary:

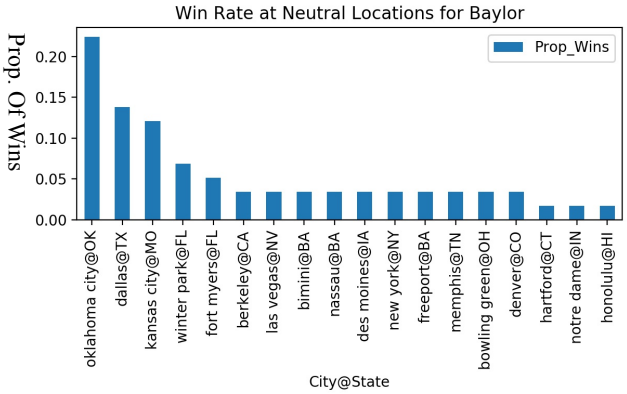
```
cleveland@OH      0.538462
fort myers@FL     0.153846
denver@CO         0.153846
dallas@TX         0.076923
seattle@WA        0.076923
Name: City_State, dtype: float64
```

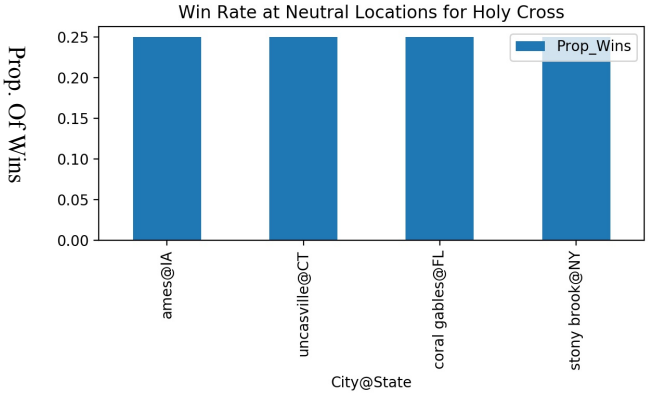
The table on the left clearly shows the proportion of games won by Akron from the 2010 through 2018 Season, across all the Neutral cities identified by a 'City@State' format do avoid errors with the same city name, but in a different state.

From the data on the left, we can see that the wins for Akron at Cleveland, Ohio is far greater than any other Neutral Location. Similarly, Purdue shows a greater number of matches won at Indianapolis, Indiana rather than other neutral location.

So clearly we can see that the proportion of wins is not equal across various neutral locations! While it is extremely hard to manually verify this trend for all colleges across we plotted histograms containing the proportion of wins in each neutral location for all colleges. We have ordered these histograms by the decreasing order of wins, and a random selection of 3 plots are shown below:







The first two scatter plots show that there proportion of wins is clearly different across neutral arenas across the U.S. However, there is some similarities between the two Histograms. For one, there seems to be some form of grouping between the neutral locations. For instance we can see that the proportion of wins at Norfolk Virginia, San Antonio Texas, and Spokane Washington happens to be the same for Stanford, and similar groups can be seen amongst other locations as well. Similar grouping of equal proportions can be seen in the proportions of wins across various neutral location for Baylor as well. However, no such trend is seen for schools such as Holy Cross, which seems to have an “ideal trend”, with wins equals scattered between neutral locations. The above graphs suggest the following:

1. For schools that show a difference of win proportions across Neutral Locations, there seems to be some form of "grouping", or equality of proportions between groups of multiple locations.
2. Secondly, they suggest that the trend observed is not not universal, and that schools do exist where the proportion of wins uniformly spread out over all neutral locations.

The differences in the distributions in the proportion of wins across neutral locations is thus conflicting across various schools and thus, to clarify the validity of this hypothesis, statistical testing is used!

So what will the null hypothesis be? On observation of the above histograms, a good null hypothesis would be to state that the proportion of wins across all the neutral locations is the same. The Alternative hypothesis would be thus, that, there is some difference in the proportion of wins across the neutral locations.

Since we are testing for equality of proportions across multiple categories, an ANOVA will **not** be appropriate in this case. However, a Chi-Squared Goodness of Fit Test from Observed Counts (Frequencies of Wins) and Expected Counts (Derived from a Uniform Distribution) will be useful in testing the null hypothesis. Under this test the null hypothesis can be equivalently stated as:

H_0 : The observed counts of wins and the expected counts of wins across neutral locations do not deviate significantly.

H_A : The observed counts of wins and the expected counts of wins across neutral locations show significant deviations.

Let us see first if the assumptions for running a χ^2 goodness-of-fit test are first founded.

Assumption I: We assume that the observations of the sample so chosen is ***independent***.

That is, knowing a team's performance in a particular location say, for example, Cleveland Ohio influence our decision about the performance of the team in Dallas, Texas. Well, if the expected counts is dependent on some parameter such as distance, which does vary between the various categories, the this assumption **could** be ***invalid***.

Thus, in order to identify through statistical means, if the proportion of wins are spatially related we must first understand the technicalities of spatial autocorrelation. While having spatially related data, *spatial autocorrelation* helps us to understand the degree to which one object is similar to nearby objects. A positive spatial autocorrelation means that similar values cluster together in a map, and a negative spatial autocorrelation is when dissimilar values cluster together. The reason why spatial autocorrelation is so important is because , if autocorrelation exists in a map, then this violates the fact that observations are independent from one another. Thus, breaking the assumption for tests such as the Chi-Squared Goodness of Fit Test.

The primary method in statistics to measure the degree of spatial correlation in statistics is the *Moran's I statistic*. The statistic ranger from -1, to 1 much like a correlation coefficient and a positive value of Moran's I statistic indicates that the values are clustered together in space, and a negative value means that the values are going to be less clustered in space. A Moran's I statistic of 0 suggests that the data is randomly spatially clustered, thus suggesting an absence of spatial autocorrelation!

The null hypothesis and alternative hypothesis used for testing spatial autocorrelation is:

H_0 : The dataset is spatially clustered as expected if the underlying spatial processes were random

H_A : The dataset is not spatially clustered as expected if the underlying spatial processes were random

Consider setting a significance level (α) of 0.05! In order to reject/validate the null hypothesis.

The values being tested in our case are the proportion of wins registered by each team at that particular location. We get the location by cross-referencing between the series and the cities dataset in order to map coordinates of each "city@state" in the index of the series to a latitude, longitude pair.

In order to compute the Moran's I statistic, we also need to compute the spatial weights matrix for the dataset. The weights matrix defines a local neighbourhood around each geographic unit. The value at each unit is compared with the weighted average of the values of its neighbours.

In our case, we use an inverse-distance between points based weighting, since we want to see whether closer points in space have higher values of the proportion of wins. However, contiguity based approaches, or other approaches such as K-Nearest-Neighbours can be used well!

Further, for this case we compute the Global spatial autocorrelation statistic in order to provide a single measure of spatial autocorrelation for an attribute in a region as a whole. Upon computing a spatial weights matrix for the dataset, we use the *ape* library in R to loading in the proportion of wins as an array as the first argument, along with the spatial weights matrix as the second argument for the Moran's I statistic. Then, we wrap our code into functions in R to create a Moran's I statistic for each school in the dataset. For instance for the institution Akron, we obtain the following results. The results for Akron suggest that the observed Moran's I statistic is a weak negative value. Thus suggesting small amount of negative autocorrelation. However, the large p-value ultimately suggests that we must retain the Null hypothesis that no "clustering" of similar/dissimilar values could be seen. The insignificant p-value, coupled with the weak negative autocorrelation pursues us to conclude that spatial autocorrelation may not be present in the proportion of wins across Neutral locations for Akron.

```
[1] "Akron.csv"
$observed
[1] -0.2208254

$expected
[1] -0.25

$sd
[1] 0.06620418

$p.value
[1] 0.6594476
```

However, the results across other Colleges do not necessarily reflect the same trend! Consider the case of Baylor, with the results described below:

```
> moran('./r_datasets/Baylor.csv')
$observed
[1] 0.1027815

$expected
[1] -0.07692308

$sd
[1] 0.08993457

$p.value
[1] 0.04569827
```

We can see quite clearly that the observed Moran -I statistic is in fact 0.1, thus suggesting a weak-*positive* spatial autocorrelation. Further, the p-value of 0.045 is under the set-significance level of 0.05. This suggests that the results of a weakly positive spatial autocorrelation are statistically significant for the proportion of wins by Baylor College across all Neutral Locations. However, since the autocorrelation is very close to zero, we can conclude that there is a statistically significant

presence of a positive autocorrelation having extremely small magnitude.

Thus, we can see that sign of the observed Moran's I statistic varies from college to college, and that the statistical significance of the Moran' I statistic also varies from one college to another.

CONCLUSIONS FROM THE TESTS FOR SPATIAL AUTOCORRELATION

1. From the tests for spatial autocorrelation, we can see that the sing of autocorrelation varies from one college to another. This suggests that the notion that the proportion of wins at neutral locations for teams **MUST** decrease with increase in distance from home location is not completely founded.

2. While the sign does vary, the extent of spatial autocorrelation seems to be almost universally vey weak, thus suggesting the the spatial autocorrelation may have no real effect, even if the signs are different owing the the minuteness in the magnitude of spatial autocorrelation.

3. However, the seeming inconsistency of the test to produce statistically significant results as seen in the case of Akron college suggests that not all of the spatial autocorrelation tests are statically valid!

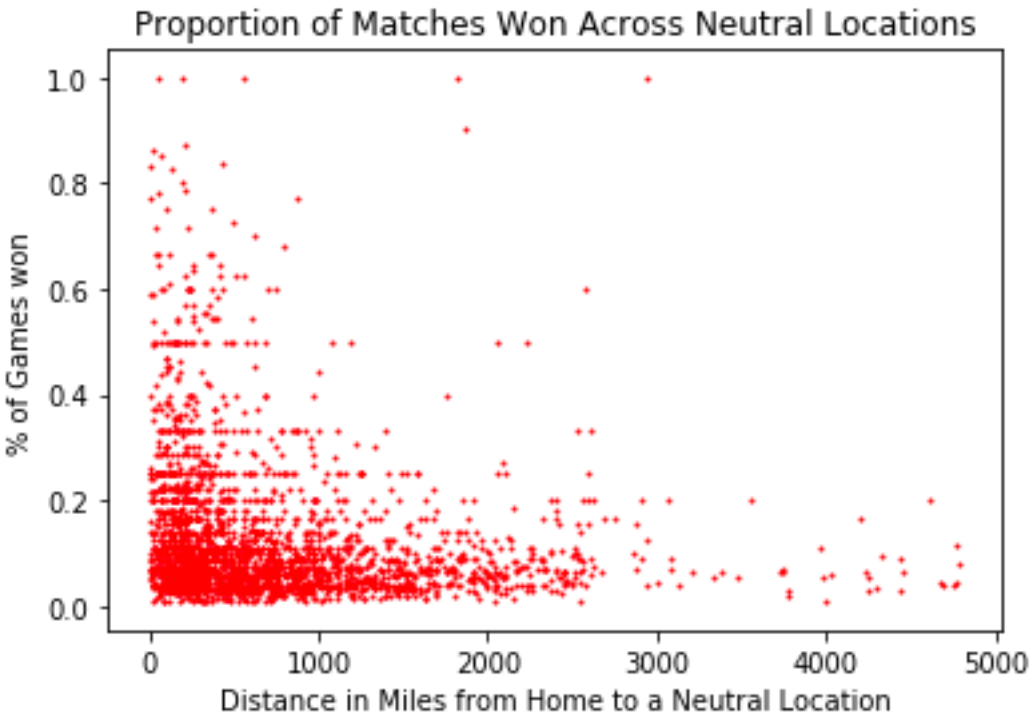
The conclusions above show that it is hard to verify that existence of spatial autocorrelation with the help of the above tests. However, a probable cause of why the tests are so hard to validate could be because of the limited above of data available. For instance, for Akron institute, we can see that we have only five points on which we must make a conclusion!

AN OVERALL DISTRIBUTION OF WINS ACROSS DISTANCE

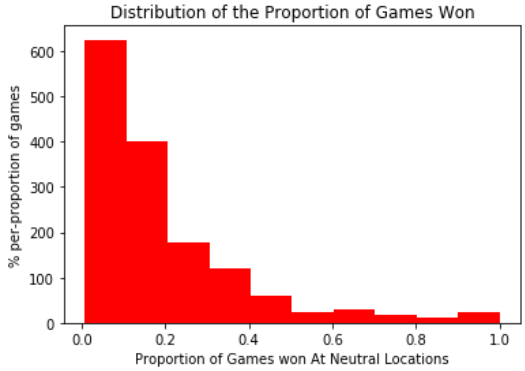
We have seen that it is hard to conclude if there is a presence of spatial correlation amongst *subsets* consisting of the proportion of wins across neutral locations for each team. However, we can look at this problem in another light!

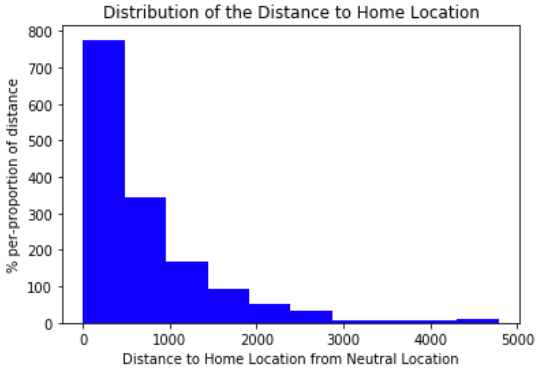
What if we take the distance between the home coordinates and the neutral location, and plotted the proportion of wins at each distance from the home location. Ideally if there were any spatial association, there must be regular trend followed with increases in distance from the home court. Further, from a logical standpoint, we would expect that the regular trend would be an inverse relationship with distance from the home court. That is, greater the distance of a natural location from the home court, greater will be the proportion of wins registered by a team here.

We gauge the diatcue between each home court and neutral location by take the coordinates for each of these from the cities.csv dataset. Then we would use the Haversine Distance as a measure of the great circle distance between the pairs of coordinates. Finally, we accumulate the data from each and every college across the U.S and obtain the following scatterplot.



The scatterplot above shows us that proportion of games does roughly share an inverse relationship with distance from the home-court to the neutral location. However, to check if any relationship exists between the two variables, that is distance and the % of games won, we can use the concept of correlation coefficients. However, on plotting histograms of the Distance in Miles to the Home Location, and % of Games Won, we observe that The distributions are not normally distributed as shown below:





As a result of this violation, we use the Spearman's Rank order correlation which is a non-parametric form of the conventional Pearson's Product Moment Correlation. The assumptions for this test are:

Assumption I: The variables are either 'ratio', 'interval' types. This assumption is satisfied in the above data since both distance, and the proportions are ratio/interval variables.

Thus, we can run a Spearman's Correlation test. An important thing to keep note is that the Spearman's correlation determines the strength and direction of the **monotonic-relationship** between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.

In Scipy, in Python the Spearman Correlation Coefficient is tested with the help of the following null hypothesis with a a significance level (α) of 0.05:

H_0 : The two sets of data are uncorrelated, that is, the correlation coefficient is 0

H_A: The two sets of data are uncorrelated, that is, the correlation coefficient is 0

On running the spearman's correlation coefficient in scipy in Python we obtain the following result:

```
In [408]: stats.spearmanr(distances, games_won)
Out[408]: SpearmanrResult(correlation=-0.31255920249601032, pvalue=3.7935198484562897e-35)
```

The results claim that the spearman correlation coefficient is -0.3125 with a p-value of $3.79*10^{-35}$ Since the length of the dataset is 46842, scipy suggests that the p-value can be interpreted as being reliable! This suggests that the variables have a moderate monotonically decreasing relationship, with high statical significance! That is, increases in distances from the home location to a neutral location, will result in decreases in the proportion of wins.

This suggests that there is in fact a statistically significant, moderate relationship between the distance to a neutral location from the home location and the proportion of wins at that location.

CONCLUSION

So with this in mind, we can see that the observations in the chi-square goodness of fit test earlier do have a possibility of being dependent after-all, and that the independence of observations **cannot be ascertained** with 100% confidence. This invalidates our approach to use the Chi-Squared Goodness of Fit Test.

However more importantly, it also suggests that the win rates of various teams at "neutral" location share a moderately monotonically decreasing relationship, with high statistical significance. This would not be the case if the locations were truly "neutral"! Thus, the unlimited indication from the segment is that there exists a moderate regular relationship between the distance of a neutral location from the home ground of a particular team, and the proportion of wins at that location, thus suggesting that neutral locations are not completely "neutral". However, the degree of this relationship is moderate at-best and is statistically highly significant. Hence, further analysis by using better methods and validation techniques must be used to investigate and this phenomenon across multiple datasets!

Feature Selection

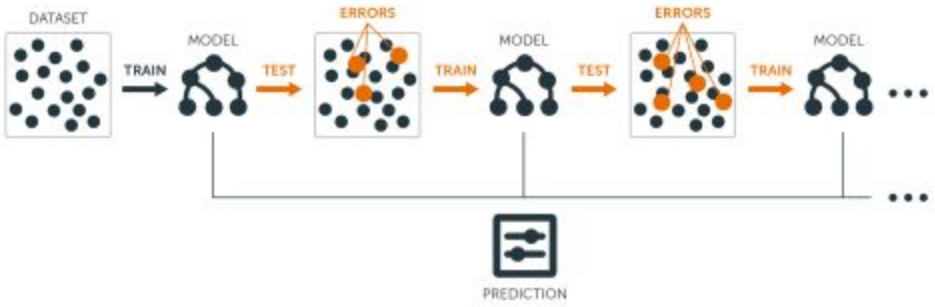
QUESTION: When predicting the winning team of a matchup between two teams, what are the factors we should use in deciding which team will win? Do some factors matter more than other factors do?

ANALYSIS:

When looking at a matchup between two teams the goal is to predict which of the two teams will win and which will lose. This is often done by comparing the stats of the two teams playing against each other. For example, we may simply choose the team with the higher seed as the predicted winner, however previous data has shown that this alone is not a good indication of who will win. The question then arises: what are the factors that we should be looking at when determining which team will be the winning team, and additionally. should we put more of a weight on a specific factor over the other factors. There are multiple machine learning models that can help us solve this problem.

A gradient boosted tree is a composite model that combines the efforts of multiple weak models to create a strong model, and each additional weak model reduces the mean squared error (MSE)(loss function) of the overall model. It is based on the idea that an ensemble of weak learners, which perform slightly better than random chance would, can become a strong learner. The prediction models are usually in the form of a decision tree. Gradient boosting involves using a method of evaluating how well the algorithm models the dataset. This method is known as the loss function. The specific

loss function used may vary significantly depending on the type of problem being solved, i.e. squared error for regression or log loss for classification.



Source: http://uc-r.github.io/gbm_regression

Additive modeling is the foundation of boosting. We take a function that maps features of an observation, x, to a scalar target value, y. It takes a bunch of simple functions and adds them together to create a more complicated function. By adding together the weaker (simpler) models, we can improve the accuracy with each iteration. In this case the observation is any given game between two teams. The features of the game are the stats of the two teams. The target y value is a one if team1 won or a zero if team2 won. Because the model takes in a single vector as input, the *difference* of the stats of the two teams will be represented in a single vector. The stats used are the number of wins in the regular season, average points scored per game in that season, the average number of points allowed to the other team per game in that season, average number of three pointers scored per game for that season, the average number of free throws made per game for that season, average number of defensive rebounds per game for that season, number of assists per game for that season, the average number of

turnovers per game for that season, the average number of steals per game for that season, the average number of fouls per game for that season, the total number of tournaments won from all years, the seed for that tournament, and lastly, the location of the game played. These stats were chosen based on knowledge of basketball. The goal of the gradient boosting model is to create a function that maps the stats vector to a one or zero representing if team1 or team2 won, so that given any two teams represented as a vector of the difference in stats, we can predict which team won. Training a model is just a matter of fitting a function through some data points. By analyzing how a gradient boosting tree predicts the y label , we will get an idea of the features it uses in classifying the vector and therefore we will know what stats should be focused on and are statistically important for choosing which team will win. A gradient boosted tree was chosen for several reasons:

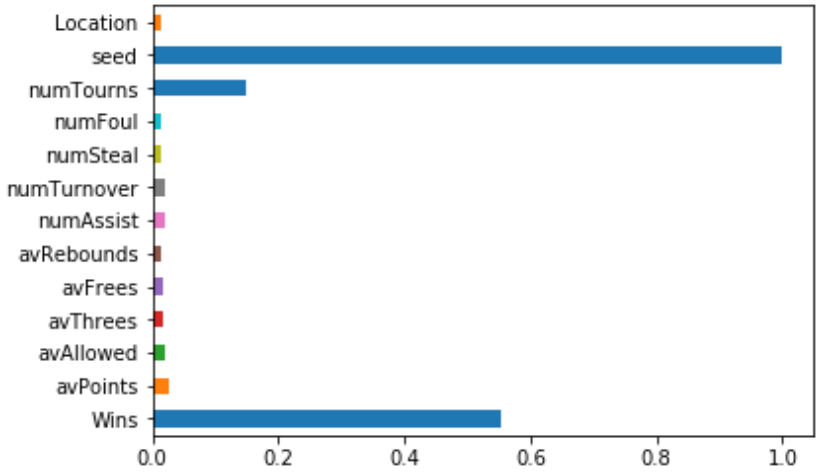
1. It can reliably extract the important features(stats) from a whole vector of stats
2. It naturally discovers nonlinear interactions between features and therefore still succeeds when features interact in nonlinear ways
3. By combining weak trees, gradient boosting trees boast a higher accuracy compared to other machine learning models because minor adjustments are made in each iteration, improving the accuracy each and avoiding overfitting.
4. Compared to nonlinear algorithms, such as a Random forest, it is not computationally expensive due to the use of many weaker models.

The gradient boosting model from sklearn was used. The model takes as input the number of boosting stages to perform. Gradient boosting is fairly robust to over fitting so a large number usually results in better performance.

After creating the training data, a vector representing the stats of the two teams and a one or zero for which won, the model was fitted to the data. Individual decision trees intrinsically perform feature selection by selecting appropriate split points. This information can be used to measure the importance of each feature. The basic idea is: the more often a feature is used in the split points of a tree the more important that feature is. This notion of importance can be extended to decision tree ensembles by simply averaging the feature importance of each tree. Using the feature selection method, we obtained the following results as output:

[3.63681652e-01 5.29588643e-03 1.67140010e-03 6.61654351e-04
 4.42280717e-04 4.18565400e-04 3.76242339e-03 3.43629100e-03
 5.70921775e-04 4.19096952e-04 1.24875912e-01 4.91103391e-01
 3.66052486e-03]

Which can be visualized from the following graph:



From the graph above we can see that the top three features that the model used in classifying were seed, numTourns, and Wins. Therefore we can conclude that in predicting which team will win, when looking at the stats of a team we should give more importance(weight) to the number of wins from their regular season, the number of tournaments they've won and most importantly the seed that they have been given.

Theory

Scatterplot

Scatterplots are extremely useful for showing general trends amongst pairs of data. Scatterplots are sometimes known as correlation plots owing to the fact that they show the extent of correlation between two variables. They are also useful for detecting non-linear relationships easily. Scatterplots are extremely versatile in the sense that the size of each point may be scaled to reflect a specific parameter. Further, the intensity of its hue may also be chawed dynamically to reflect another variable. Finally, scatterplots may also be used for geographic, and geo-spatial analysis for the identification of points across maps.

Barchart

Bar-charts are probably one of the most common representation of the of the frequencies of values across various categories. The domain of categories in a bar-chart is discrete, and the width of each bar in the bar-plot is always equal. The height of each bar corresponds to the frequency of observed values in that particular category. The orientation of bar plots can be either vertical or horizontal. That is the categorical values can either be on the x-axis or the y-axis. However, a general practice especially while visualising bar-plots is that, if there exists some natural order for the categories, the bars must be ordered in that specific order. If no set-order exists, a good practice is to either order the bars by ascending frequency of values, or descending frequency of values in order to aid quicker judgements.

Histograms

Histograms are probably one of the most common representation of numeric data today.Histograms are used exclusively to depict relationship between numerical (quantitative data) In fact a histogram can be thought of as one of the simplest non-parametric estimation techniques. The histogram essentially groups data with the help of non-overlapping segments termed as bins. The data is then grouped into each bins. One important thing to note in a histogram is that the bins *need not be of equal width*. Thus, the height corresponding to each bin in a histogram is not always proportional to the frequency of observed values in a given bin. However, in a special case when the bins happen to be of equal widths, the height of a histogram does happen to be
Proportional to the frequency of values observed in each bin. However, more generally the area of a bin is proportional to the to the frequency of the values in that particular bin. Finally, a histogram can also be normalised, and in this particular case it proves an account of the proportion of values which fall into each bin category, and in this specific case the sum of heights happen to be 1, since the sum of portion of values must equal 1, on account of the applied normalisation.

While the histogram is a non-parametric estimation technique, consider the histogram to be a parametric model, parametrised by the bin heights h1….hk, given fixed bins b0…..bk. Assume that the bin widths are equal! In that case we can assume that the associated probability mass function is

$$f_X(x) = h_{\beta(x)}$$

where $\beta(x)$ is the index of the bin containing x.

Hence, the likelihood function for an MLE based estimate would be :

$$L(h_1, \ldots, h_K) = \prod_{k=1}^{K} h_k^{n_k}$$

And we would like to maximise this likelihood function subject to the constraint,

$$\int f_X(x)\, dx = 1.$$

Using Lagrange multipliers, this estimate provides the height at a particular bin (hk) to be

$$\hat{h}_k = \frac{n_k}{n\Delta_k}$$

Where the Δk is essentially the bin width!

This suggests that if we hypothetically know that the number of bins are fixed, and are of equal widths, we can estimate the height of the height of the histogram just from the number of data points, the bin width and the total number of data points through the process of maximum likelihood estimation. While not used in this investigation, the above result could be used to yield powerful results elsewhere!

Quantile-Quantile (QQ) Plots

Quantile-quantile plots can be used in this investigation because they can be used for any distribution. They compare two datasets but pairing their sample quantiles. To find the sample quantiles, the data is ordered in ascending order. The p-th quantile of a random variable X is a number q where P(Xq)p and P(Xq)(1-p). The x(k)is at the k_{n+1th} sample quantile. By plotting theoretical and sample quantities, any departure from a straight line is a sign of departures from normality. Identical distributions will have an intercept of 0 and a slope of 1. N intercept not at 0 means that there are shifts in both distributions, and a non unit hints at a scale change. The plot can also be linear if the means or standard deviation of the distributions are different but the shape is similar. Other departures of normality are indicated by an upward curve (long right tail), downward curve (long left tail), stripes (granularity), and curved middle (bi-modality).

Kolmogorov-Smirnov Test

The KS test (Kolmogorov-Smirnov test) is a nonparametric method that can be used to compare two samples. It is useful because it is sensitive to differences and location and shape of the one-dimensional probability distribution for the two samples, which is important when finding the difference in weight between babies born to mothers who smoked during pregnancy and those who did not.

In this investigation, a variation of the Kolmogorov Smirnov Test known as the K.S-2 Sample Test is used. According to the Statistical Engineering Division of the National Institute of Standards and Technology (NIST), the K.S 2 Sample test can be defined as taking the absolute difference between the 2 empirical distribution function of ordered iid random variables. Mathematically, this may be written as: $D=|E_1(i)-E_2(i)|$ where E_1 and E_2 are the empirical distribution functions for the two samples, and is computed at each point (i) in the sample.

The NIST also provides the following definition for the K.S two sample test in terms of the null and alternative hypotheses involved[10] :

H_0:	The two samples come from a common distribution.
H_a:	The two samples do not come from a common distribution.
Test Statistic:	The Kolmogorov-Smirnov two sample test statistic is defined as $$D = \|E_1(i) - E_2(i)\|$$ where E_1 and E_2 are the empirical distribution functions for the two samples.
Significance Level:	α
Critical Region:	The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value obtained from a table. There are several variations of these tables in the literature that use somewhat different scalings for the K-S test statistic and critical regions. These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated.

Kernel Density Estimation (KDE)

Kernel Density Estimation[13] is a non-parametric estimation technique to estimate the probability distribution of a continuous random variable that independent and identically distributed (iid) from a sample drawn from a population of data. KDE uses a smoothing function (typically a normal distribution) to estimate the P.M.F of the data using the following formula:

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

Where, $f_h(x)$ is the kernel density function (or the theorised P.M.F), K is the smoothing function. In our investigation, we have used the gaussian (normal) distribution as the smoothing function. The only constraint is that the Function K must be a non-negative function. Further, h is the bin width, which must be greater than 0. In general lower the bin width, more granular the PMF, and higher the bin width more coarse the PMF is going to be. In our investigation, we have used the auto-binning feature which uses an in built algorithm to decide on the most optimal bin width.

Skew and Kurtosis

Two statistics that check for normality are skewness and kurtosis.

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 \text{ and Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

The *skewness coefficient* is the average of the third power of the standardized data and how symmetric the data is, while the *kurtosis coefficient* is the average of the fourth power of the standardised data and tells us how pronounced the peak of the distribution is. If the distribution is symmetric, the skewness coefficient will be 0. If the distribution is normal, the kurtosis coefficient will be 3. Skewness and Kurtosis were used as measures of normality for each infant birth weight distribution. These statistics were calculated in order to check if both distributions satisfy the condition of normality in order to run the ANOVA.

GOODNESS OF FIT TEST USING CHI-SQUARE

In essence, the Chi-Square goodness of fit test is a non-parametric test that is used to find out how the observed value of a given phenomena is significantly different from the expected value. In Chi-Square goodness of fit test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution. Chi-Square goodness of fit test determines how well theoretical distribution In Chi-Square goodness of fit test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval. In order to validate our claims we then perform a hypothesis by considering the following null and alternative hypothesis tests. An important assumption is that the observations are independent of each other!

A. **Null hypothesis:** In Chi-Square goodness of fit test, the null hypothesis assumes that there is no significant difference between the observed and the expected value.

B. **Alternative hypothesis:** In Chi-Square goodness of fit test, the alternative hypothesis assumes that there is a significant difference between the observed and the expected value.

Whether the differences are significant or not depends on the coupled value of χ^2 test-statistic for the dataset and the number of degrees of freedom. In order to compute the χ^2 test-statistic we follow the following formula: $\chi^2 = \left[\frac{(O-E)^2}{E}\right]$ Where, χ^2 is the Chi-Square goodness of fit test statistic, O is the observed value(s) and E is the expected value. The χ^2 values are the summed up over all observations, to produce the complete χ^2 test-statistic for the dataset. Generally, the test-statistic is computed by following the following procedure:

$$\sum_{j=1}^{m} \frac{(\text{jth sample count} - \text{jth Expected count})^2}{\text{jth Expected count}} = \sum_{j=1}^{m} \frac{(N_j - \mu_j)^2}{\mu_j}.$$

Further it is important to note that the general convention while computing the χ^2 test-statistic is to keep the expected counts of each row of the table to be greater than or equal to five! To do this, the usual approach is to combine two or more rows of the goodness of fit table so that the expected value of the combined row reaches a value that is greater than or equal to five. However, additional testing through simulation can only suggest if this approach is revenant to the context of an investigative study.

The χ^2 distribution is going to be a continuous distribution on the positive real line and the density has a long right tail. As the degrees of freedom increase it starts to look symmetric and looks similar to the normal distribution. Using the χ^2 distribution and the degrees of freedom, we can now compute the associated p-value to obtain a goodness of fit. If the p-value is small, then there is a reason to doubt the fit of the distribution.

In case the p-value is extremely small a residual plot can help determine where the lack of fit occurs. For each category, we can plot the standardised residuals by using the formula

$$\frac{\text{sample count} - \text{Expected count}}{\sqrt{\text{Expectedcount}}} = \frac{N_j - \mu_j}{\sqrt{\mu_j}}.$$

The denominator transforms residuals in order to give them approximately equal variance, and the Square root makes it easier to make meaningful comparisons across categories. It is important to note here that the sum of residuals is going to zero, whereas the sum of standardised residuals is never zero. In general value of standardised residuals larger than 3 indicate a lack of fit.

The Simple Linear Regression Model

The simple linear model says that the expectation $E(Y|x)$ of a random response Y at a known x satisfies the relation

$$E(Y|x)=a+bx.$$

An extension of this model is the *Gauss measurement model* which says that

$$Y = a + bx + E,$$

Where *E* represents the measurement errors. *E*s have mean 0, are uncorrelated, and constant variance (σ2).

If our observations are of the form (x_1, y_1), …. , (x_n, y_n), then using the method of least squares we can estimate *a* and *b.* So we get

$$\hat{a} = ((\sum x_i^2)(\sum y_i) - (x_i)(\sum x_i y_i)) \div (n\sum x_i^2 - (\sum x_i)^2)$$

$$\hat{b} = (n\sum x_i y_i - (\sum x_i)(\sum y_i)) \div (n\sum x_i^2 - (\sum x_i)^2)$$

Based on this the residual sum of squares has the expectation

$$E(\sum[y_i - (\hat{a} + bx_i)]^2) = \widehat{(n-2)}\sigma^2.$$

The residual sum of squares can therefore provide an estimate of the variance in the Gauss measurement model.

Residuals and the Best Line

Residuals are the leftovers from the model fit: Data = Fit + Residual. It is the difference between the observed y_i and the predicted $\hat{y}$ under the equation $e_i = y_i - \hat{y}_i$. A measure for the best line has a small residuals which we commonly get through minimizing the sum of squared residuals (least squares) $e_1^2 + e_2^2 + ... + e_n^2$. Least squares is favored because it is the most used, easier to compute by hand/software, and a residual twice as large as another is usually more than twice as bad.

The Least Squares Line

$\hat{y} = \beta_0 + \beta_1 x$, where $\hat{y}$ is the predicted y, β_0 is the intercept, β_1 is the slope, and x is the explanatory variable. A point estimate of the two parameters represented by b_0 for the intercept and b_1 for the intercept.

The slope of the regression line has the equation $b_1 = \frac{s_y}{s_x}R$, and can be interpreted as for each unit in x, y is expected to increase/decrease on average by the slope, all else held constant. It is important to note that the interpretation of the slope and intercept are not casual, unless the study is a randomized controlled experiment.

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact that a regression line always passes through $b_0 = \overline{y} - b_1\overline{x}$, and can be interpreted as when x=0, y is expected to equal the intercept, all else held constant. But sometimes the intercept is of no interest, not reliable, and not useful if for example the predicted value of the intercept is far from the bulk of the data.

Prediction

The linear model is useful because we can predict the value of the response variable for a given value of the explanatory variable using a process called prediction. This involves plugging in the value of x into the linear model equation. There is uncertainty associated with the predicted value. Applying a model estimate to values outside of the realm of the original data is called extrapolation. The intercept can be an extrapolation especially if the predicted value of the intercept is far from the bulk of the data.

Conditions for the Least Squares Line

The first condition is linearity. The relationship between the explanatory and the response variable should be linear. We can check linearity by using a scatterplot of the data or a residual plot. The second condition is nearly normal residuals, meaning that the residuals should be nearly normal. If there are unusual observations that don't follow the trend of the rest of the data, this condition may not be satisfied. We can check this using a histogram or normal probability plot of residuals. The third condition is constant variability. The variability of points around the least squares line should be roughly constant, thereby implying that the variability of residuals around the 0 line should be roughly constant as well (homoscedasticity). We can check this using a histogram or normal probability plot of residuals.

Transformations

The relationship between the predictor and response variables may not necessarily be linear at first. Transformation can be used to put this relationship in a linear form. In this investigation, for instance, we used a log transformation for both the predictor and response variable!

R-SQUARED VALUES*

The R^2 values between 2 variables essentially gives us how good a linear relationship is shared between the two variables. However, technically speaking the R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. However, in this context, since we know that the two variables share a linear relationship with each other, we can use a R^2 value to explain how much of the variance in y explained with the help of X. In this context, an R^2 value of 1 suggests that the relationship is almost certainly linear between X and Y, and suggests that the two values were drawn from the same distribution! However, in the conventional definition of R^2 values a value of 0% indicates that the model explains none of the variability of the response data around its mean, whereas a value of 100% indicates that the model explains all the variability of the response data around its mean. In general, higher the R^2 value, the greater a linear relationship is shared between the x and y variables.

*(material sourced from http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.)

Haversine Distance

The haversine formula is used to calculate the rest circle distance between a pair of coordinates on the spherical surface of the earth. The haversine formula gives the “as the crow flies” distance between a pair of coordinates. The Haversine Distance is particularly useful in our case cause it remains well-conditioned for computation even at extremely small distances. The formula is most commonly written in the form:

$$d = 2r \arcsin\left(\sqrt{\operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)}\right)$$
$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Where *d* is the distance in the units of the chosen *r,* and *r* is the radius of the sphere (which is the Earth in our case). Finally the coordinates are written as (φ_1, λ_1), which is basically the latitude and longitude of the first point and (φ_2, λ_2) which is the latitude and longitude of the second point! It must be notes that the radius of the Earth, *r,* is approximately 3,958.8 miles.

Spatial Autocorrelation

According to a paper by Nilupa et. al. from Purdue University *The phenomenon due to which observation made at closer locations may share a regular relationship with each other rather than observations that are far apart is termed as spatial autocorrelation.* According to the paper, Spatial autocorrelation measures the correlation of a variable with itself through space. Further, Spatial autocorrelation can be positive or negative. Positive spatial autocorrelation occurs when similar values occur near one another. Negative spatial autocorrelation occurs when dissimilar values occur near one another. In order to determine the existence of spatial correlation the first step is to determine the spatial weights matrix. The spatial weights matrix represents what is meant by two observation being close together, in other words *a distance based metric is required*. The distances (or the equivalent metric) are presented in the weights matrix, which defined the locations where the measurements are made. In general the matrix is going to be a *n by n* matrix for n datapoint with zeros on the diagonal. The weights in turn can be specified in many different ways. The paper by Nilupa et. al, specifies four broad ways of quantifying weights, namely:

1. The weight for any two different locations is a constant.
2. All observations within a specified distance have a fixed weight.
3. K nearest neighbours have a fixed weight, and all others are zero.
4. Weight is proportional to inverse distance, inverse distance squared, or inverse distance up to a specified distance

In this investigation we have used the fourth approach to calculate the spatial weights, by considering the inverse distance.

In order to measure the degree of spatial autocorrelation, a statistic known as the Moran's I Statistic is used. According to a paper by S. Oliveau of the UMR Géographie-cités, Paris,

Moran introduced in 1950 the first measure of spatial autocorrelation in order to study stochastic phenomena, which are distributed in space in two or more dimensions. Moran's index has been subsequently used in almost all studies employing spatial autocorrelation. Moran's I is used to estimate the strength of this correlation between observations as a function of the distance separating them (correlograms).

It shares many similarities with Pearson's correlation coefficient: its numerator is a covariance, while its denominator is the sample variance. And like a correlation coefficient the values of Moran's I range from +1 meaning strong positive spatial autocorrelation, to 0 meaning a random pattern to -1 indicating strong negative spatial autocorrelation –although negative autocorrelation is extremely unusual in social sciences.

The precise definition of Moran's I is given below for a spatialized variable z_i at location i.

$$I = \frac{\sum_{i,j} W_{ij}(z_i - \bar{z}).(z_j - \bar{z})}{n} / \sigma^2(z)$$

Where σ^2 is the sample variance

Thus, the Moran's I statistic helps us to gauge the extent of spatial autocorrelation with the help of a single statistic!

Ultimately, it is very important and crucial to note that Commonly used statistical approaches often assume that the measured outcomes are independent of each other. In spatial data, it is often the case that some or all outcome measures exhibit spatial autocorrelation. This occurs when the relative outcomes of two points is related to their distance. When analysing spatial data, it is important to check for autocorrelation. If there is no evidence of spatial autocorrelation, then proceeding with a standard approach is acceptable. However, if there is evidence of spatial autocorrelation, then one of the underlying assumptions of your analysis may be violated and your results may not be valid.

The Z- and Students *t* Test for Proportions

The Z-test for proportions, is a statistical test used to determine whether two population proportions are different when the variances are known and the sample size is large. The test statistic is assumed to be normally distributed. Consider $\hat{p}_1$ to be the first proportion and $\hat{p}_2$ to be the second proportion in the sample. The Z statistic to be used is calculated by using the formula: where $\hat{p}$ is the pooled proportion computed by using the formula:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}} \text{ where } \hat{p} = \frac{p_1 \cdot n_1 + p_2 \cdot n_2}{n_1 + n_2} \text{ where } n_1 \text{ is the number of items in}$$

sample 1 and n_2 is the number of items in sample 2. The calculated values of the statistic can then be used to estimate p-value, and then validate null/alternative hypothesis. In a Z test, it is a general rule of thumb that the sample sizes, n_1 and n_2 are greater than or equal to 30, or alternatively the population is normally distributed. Further more, Independence of each observation is also implicitly assumed. Finally, the population standard deviation σ is assumed to be known as well. In case the assumptions regarding the number of samples, falls shorts, a test involving looking up values from a student's t-distribution is often used in place of the z-score.

The Spearman's Rank Order Correlation

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, (ρ, also signified by r_s) measures the strength and direction of association between two ranked variables. The main assumptions of the test are that the variable types are either an interval type, that is the values of this data type represent ordered values which happen to have the same difference, but may or may not have a true "zero". The main operations which can be performed on interval data are addition, and subtraction. However, multiplication and division are not possible. Ratio values on the other hand can be multiplied and divided! They also have a well defined true zero. Finally their values also represent ordered values which happen to have the same difference. Finally, the Spearman's rank-order correlation produces a ρ (correlation coefficient) which quantifies the degree of a monotonic relationship shared between the two variables. Negative values of the correlation coefficient indicate negative monotonic relationships and positive values of the correlation coefficients indicate positive monotonic relationships. The spearman's rank-order correlation is most frequently used when the assumption under the conventional Pearson's correlation such as the normality of variables is not satisfied.

Gradient Boosting Tree

Boosting is a machine learning technique that combines multiple weaker models into a single composite model. The newer models are sequentially added to the previous models one at a time. The idea is that, as we introduce more simple models, the overall model becomes a stronger and stronger predictor. In this investigation, a gradient boosting tree was used to extract the most important features used in classifying a team as the winning or losing team. Given an x, in this case a vector representing the stats of two teams, we'd like to learn the scalar target value y for a bunch of (x,y) pairs. The target value y represents the winning team. In the case of a boosting algorithm, the target value y is expressed as the sum of multiple weak models or functions.

Boosting is able to improve the overall model performance by constructing and then adding the weak models one after the other. The boosting algorithm can be thought of as a greedy algorithm because the newer functions or models do not alter the previous functions. Instead small changes to the model are added to the overall model.

$$F_0(\mathbf{x}) = f_0(\mathbf{x})$$
$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \Delta_m(\mathbf{x})$$

After one iteration is complete, the residual is computed as the actual minus the predicted. The new weak model is then trained on the residual vector data. If the new model is still not satisfactory then another model is added. Once we reach a stage that residuals do not have any pattern that could be modeled, we can stop modeling residuals thus we are minimizing our loss function. The mean squared error is the most common loss function used in gradient boosting.

$$y_i^p = y_i^p + \alpha * \delta \sum (y_i - y_i^p)^2 / \delta y_i^p$$

which becomes, $y_i^p = y_i^p - \alpha * 2 * \sum (y_i - y_i^p)$

where, α is learning rate and $\sum (y_i - y_i^p)$ is sum of residuals

So, we are basically updating the predictions such that the sum of our residuals is close to 0 (or minimum) and predicted values are sufficiently close to actual values.

So, the intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better. Once we reach a stage that residuals do not have any pattern that could be modeled, we can stop modeling residuals**.** *In the end, we combine all the predictors by giving some weights to each predictor.*

CONCLUSION

In this study, historical and current data on the NCAA Women's Basketball regular season and tournament games were analyzed in order to research what features can be used as the principal factors in determining the outcome of games. These features included both in-game features such as points scored and also external features such as day and location. In particular, due to the frequently observed phenomenon known as the *home court advantage*, the study attempted to test whether the difference in performance dependent on location is statistically significant. The nature of the relationship between location and team performance cannot be fully established since not all teams had statistically significant differences between their performance at home and away courts. Furthermore, the gradient boosted tree model did not even consider using location as its top three features. Nevertheless, the study paves the way for further research into other features that could potentially be used as the determining factor in predicting match outcomes.

References

1. Oliveau, Sébastien. “Spatial correlation and demography . Exploring India ’ s demographic patterns.” (2005).

2 . Morris Tracy L. & Bokhari Faryal H., 2012.
"The Dreaded Middle Seeds - Are They the Worst Seeds in the NCAA Basketball Tournament?," Journal of Quantitative Analysis in Sports, De Gruyter, vol. 8(2), pages 1-13, June.
<https://ideas.repec.org/a/bpj/jqsprt/v8y2012i2n2.html>

3. Grundy, Pamela, and Susan Shackelford. Shattering the glass: The remarkable history of women's basketball. UNC Press Books, 2017.

4. David A. Harville & Michael H. Smith (1994) The Home-Court Advantage: How Large is it, and Does it Vary from Team to Team?, The American
Statistician, 48:1, 22-28, DOI: 10.1080/00031305.1994.10476013

5. Bhandari, Inderpal, et al. "Advanced scout: Data mining and knowledge discovery in NBA data." Data Mining and Knowledge Discovery 1.1 (1997): 121-125.

6. Tukey, J. W. (1977), Exploratory Data Analysis , Addison-Wesley .

7. Cliff, Andrew D., and Keith Ord. “Spatial Autocorrelation: A Review of Existing and New Measures with Applications.” Economic Geography, vol. 46, 1970, pp. 269–292. JSTOR, www.jstor.org/stable/143144.\

8. The National Institute of Standards and Technology. “KOLMOGOROV SMIRNOV TWO SAMPLE.” 1.3.5.15. Chi-Square Goodness-of-Fit Test, https://itl.nist.gov/div898/software/dataplot/refman1/auxillar/ks2samp.htm

9. Myers, Leann, and Maria J. Sirois. "Spearman correlation coefficients, differences between." Encyclopedia of statistical sciences 12 (2004).

10. Encyclopedia Britannica. (2019). National Collegiate Athletic Association | American organization . [online] Available at: https://www.britannica.com/topic/National-Collegiate-Athletic-Association [Accessed 13 Mar. 2019].

11. Gauthier, Thomas D. "Detecting trends using Spearman's rank correlation coefficient." Environmental forensics 2.4 (2001): 359-362.

12. Ord, J. Keith. "Spatial Autocorrelation: A Statistician's Reflections." Perspectives on Spatial Data Analysis. Springer, Berlin, Heidelberg, 2010. 165-180.

13. Legendre, Pierre. "Spatial autocorrelation: trouble or new paradigm?." Ecology 74.6 (1993): 1659-1673.

14. Rodney K. Smith, The National Collegiate Athletic Association's Death Penalty: How Educators Punish Themselves and Others, 62 IND. L.J. 985, 988-89 (1987) [hereinafter Smith, Death Penalty]; Rodney K. Smith, Little Ado About Something: Playing Games With the Reform of Big-Time Athletics, 20 CAP. U. L. Rev. 567, 569-70 (1991) [hereinafter Smith, Little Ado].

15. Divisional Differences and the History of Multi-division Classification . (2013). NCAA.org - The Official Site of the NCAA . Retrieved 14 March 2019, from http://www.ncaa.org/about/who-we-are/membership/divisional-differences-and-history-multidivision-classification

16. DI Women's Basketball Championship History | NCAA.com . (2019). Ncaa.com .

17. GamesCricketRugbyCFL, &., League, N., Softball, N., Sports, O., FB, R., & BB, R. et al. (2019). Pac-12 Conference Standings - Women's College Basketball - ESPN . ESPN.com .

18. GamesCricketRugbyCFL, &., League, N., Softball, N., Sports, O., FB, R., & BB, R. et al. (2019). Nothing but NET: NCAA boots RPI for evaluation . ESPN.com .

19. *2018 Division I Women's Basketball Official Bracket | NCAA.com . (2019). Ncaa.com .*

20. Entine, Oliver A., and Dylan S. Small. "The role of rest in the NBA home-court advantage." Journal of Quantitative Analysis in Sports 4.2 (2008).

21. "ANOVA – Simple Introduction." SPSS Tutorials, www.spss-tutorials.com/anova-what-is-it/. https://www.spss-tutorials.com/anova-what-is-it/

22. "Linear Regression." Categorical Data, Yale University, http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.

23. *Breheny, Patrick. Kernel Density Estimation, University of Kentucky,* https://web.as.uky.edu/statistics/users/pbreheny/621/F10/notes/10-28.pdf

24. Data.world. (2019). Historical NCAA Forecasts - dataset by fivethirtyeight. [online] Available at: https://data.world/fivethirtyeight/historical-ncaa-forecasts [Accessed 26 Mar. 2019].

25. GitHub. (2019). cphalpert/census-regions. [online] Available at: https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv [Accessed 26 Mar. 2019].

26. Gist. (2019). *US states in JSON form*. [online] Available at: https://gist.github.com/mshafrir/2646763 [Accessed 26 Mar. 2019].