**NCAA data**

Each season there are thousands of college basketball games played between NCAA® Division I men's college teams, culminating in March Madness®, the 68-team national championship tournament that starts in the middle of March. We have provided a large amount of historical data about college basketball games and teams, going back many years. Armed with this historical data, you can explore it and develop your own distinctive ways of predicting March Madness® game outcomes. You can even evaluate and compare different approaches by seeing which of them would have done best at predicting tournament games from the past.

If you are unfamiliar with the format and intricacies of the NCAA® tournament, we encourage reading the [wikipedia page](http://en.wikipedia.org/wiki/NCAA_Men%27s_Division_I_Basketball_Championship) before diving into the data.  The data description and schema may seem daunting at first, but is not as complicated as it appears.

As a reminder, you are encouraged to incorporate your own sources of data. We have provided extensive historical data to jump-start the modeling process, and this data is self-consistent (for instance, dates and team ID's are always treated the same way). Nevertheless, you may also be able to make productive use of external data. If you head down this path, please be forewarned that many sources have their own distinctive way of identifying the names of teams, and this can make it challenging to link up with our data. The TeamSpellings file, which is listed in the bottom section below, may help you map external team references into our own Team ID structure, and you may also need to understand exactly how dates work in our data.

Please also note that we have standardized the spelling of column names and some filenames, so if you are re-using code from previous instances of this contest, you may need to adjust for this. For example, we are universally referencing Team ID columns with a spelling of "TeamID" rather than "team\_id". And this year the seeds file is NCAATourneySeeds.csv rather than TourneySeeds.csv

We extend our gratitude to [Kenneth Massey](http://www.masseyratings.com/) for providing much of the historical data.

Special Acknowledgment to Jeff Sonas of Sonas Consulting for his support in assembling the dataset for this competition.

What to predict

Stage 1 - You should submit predicted probabilities for every possible matchup in the past 4 NCAA® tournaments (2014-2017).

Stage 2 - You should submit predicted probabilities for every possible matchup before the 2018 tournament begins.

Refer to the Timeline page for specific dates. In both stages, the sample submission will tell you which games to predict.

File descriptions

Below we describe the format and fields of the contest data files. The data will likely be refreshed once in late February while Stage 1 of the competition is running. At the start of Stage 2, we will provide updates to these files to incorporate data from the current season.

Data Section 1 - The Basics

This section provides everything you need to build a simple prediction model and submit predictions.

* Team ID's and Team Names
* Tournament seeds since 1984-85 season
* Final scores of all regular season, conference tournament, and NCAA® tournament games since 1984-85 season
* Season-level details including dates and region names
* Example submission file for stage 1

Special note about "Season" numbers: the college basketball season lasts from early November until the national championship tournament that starts in the middle of March. For instance, this year the first men’s Division I games were played on November 10th, 2017 and the men’s national championship game will be played on April 2nd, 2018. Because a basketball season spans two calendar years like this, it can be confusing to refer to the year of the season. By convention, when we identify a particular season, we will reference the year that the season ends in, not the year that it starts in. So for instance, the current season will be identified in our data as the 2018 season, not the 2017 season or the 2017-18 season or the 2017-2018 season, though you may see any of these in everyday use outside of our data.

Data Section 1 file: Teams.csv

This file identifies the different college teams present in the dataset. You will not see games present for all teams in all seasons, because the games listing is only for matchups where both teams are Division-I teams. There are 351 teams currently in Division-I, and an overall total of 364 teams in our team listing (each year, some teams might start being Division-I programs, and others might stop being Division-I programs). Each team has a 4 digit id number.

* TeamID - a 4 digit id number, from 1000-1999, uniquely identifying each NCAA® men's team. A school's TeamID does not change from one year to the next, so for instance the Duke men's TeamID is 1181 for all seasons. To avoid 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.
* 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, e.g., Texas A&M, St Mary's CA, TAM C. Christi, and Bethune-Cookman.
* FirstD1Season - the first season in our dataset that the school was a Division-I school. For instance, FL Gulf Coast (famously) was not a Division-I school until the 2008 season, despite their two wins just five years later in the 2013 NCAA® tourney. Of course, many schools were Division-I far earlier than 1985, but since we don't have any data included prior to 1985, all such teams are listed with a FirstD1Season of 1985.
* LastD1Season - the last season in our dataset that the school was a Division-I school. For any teams that are currently Division-I, they will be listed with LastD1Season=2018, and you can confirm there are 351 such teams. It has been a few years since any teams stopped being Division-I; the last was Centenary whose final Division-I year was 2011.

Data Section 1 file: Seasons.csv

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

* Season - indicates the year in which the tournament was played
* DayZero - tells you the date corresponding to daynum=0 during that season. All game dates have been aligned upon a common scale so that the championship game of the final tournament is on daynum=154. Working backward, the national semifinals are always on daynum=152, the "play-in" games are on days 134/135, Selection Sunday is on day 132, and so on. All game data includes the day number in order to make it easier to perform date calculations. If you really want to know the exact date a game was played on, you can combine the game's "daynum" with the season's "dayzero". For instance, since day zero during the 2011-2012 season was 10/31/2011, if we know that the earliest regular season games that year were played on daynum=7, they were therefore played on 11/07/2011.
* 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. This allows us to identify the regions and brackets in a standardized way in other files. For instance, during the 2012 tournament, the four regions were East, Midwest, South, and West. Being the first alphabetically, East becomes W. Since the East regional champion (Ohio State) played against the Midwest regional champion (Kansas) in the national semifinals, that makes Midwest be region X. For the other two (South and West), since South comes first alphabetically, that makes South Y and therefore West is Z. So for that season, the W/X/Y/Z are East,Midwest,South,West. And so for instance, Ohio State, the #2 seed in the East, is listed in the NCAATourneySeeds file with a seed of W02, meaning they were the #2 seed in the W region (the East region). We will not know the final W/X/Y/Z designations until Selection Sunday, because the national semifinal pairings in the Final Four will depend upon the overall ranks of the four #1 seeds.

Data Section 1 file: NCAATourneySeeds.csv

This file identifies the seeds for all teams in each NCAA® tournament, for all seasons of historical data. Thus, there are between 64-68 rows for each year, depending on whether there were any play-in games and how many there were. In recent years the structure has settled at 68 total teams, with eight "play-in" games leading to the final field of 64 teams entering Round 1 on Thursday of the first week. We will not know the seeds of the respective tournament teams, or even exactly which 68 teams it will be, until Selection Sunday on March 11, 2018.

* Season - the year that the tournament was played in
* Seed - this is a 3/4-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 (either 01, 02, ..., 15, or 16) tells you the seed within the region. For play-in teams, there is a fourth character (a or b) to further distinguish the seeds, since teams that face each other in the play-in games will have seeds with the same first three characters. For example, the first record in the file is seed W01, which means we are looking at the #1 seed in the W region (which we can see from the "Seasons.csv" file was the East region).
* TeamID - this identifies the id number of the team, as specified in the Teams.csv file

Data Section 1 file: RegularSeasonCompactResults.csv

This file identifies the game-by-game results for many seasons of historical data, starting with the 1985 season (the first year the NCAA® had a 64-team tournament). 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=132 is Selection Sunday, and there are always a few conference tournament finals actually played early in the day on Selection Sunday itself). Thus a game played on or before Selection Sunday will show up here whether it was a pre-season tournament, a non-conference game, a regular conference game, a conference tournament game, or whatever.

* Season - this is the year of the associated entry in Seasons.csv (the year in which the final tournament occurs)
* 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 "Seasons.csv" file. For example, the first game in the file was DayNum=20. Combined with the fact from the "Seasons.csv" file that day zero was 10/29/1984 that year, this means the first game was played 20 days later, or 11/18/1984. There are no teams that ever played more than one game on a given date, so you can use this fact if you need a unique key (combining Season and DayNum and WTeamID). In order to accomplish this uniqueness, we had to adjust one game's date. In March 2008, the SEC postseason tournament had to reschedule one game (Georgia-Kentucky) to a subsequent day, so Georgia had to actually play two games on the same day. In order to enforce this uniqueness, we moved the game date for the Georgia-Kentucky game back to its original scheduled date.
* WTeamID - this identifies the id number of the team that won the game, as listed in the "Teams.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.
* WScore - this identifies the number of points scored by the winning team.
* LTeamID - this identifies the id number of the team that lost the game.
* 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.
* NumOT - this indicates the number of overtime periods in the game, an integer 0 or higher.
* 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". Sometimes it is unclear whether the site should be considered neutral, since it is near one team's home court, or even on their court during a tournament, but for this determination we have simply used the Kenneth Massey data in its current state, where the "@" sign is either listed with the winning team, the losing team, or neither team. If you would like to investigate this factor more closely, we invite you to explore Data Section 3, which provides the city that each game was played in, irrespective of whether it was considered to be a neutral site.

Data Section 1 file: NCAATourneyCompactResults.csv

This file identifies the game-by-game NCAA® tournament results for all seasons of historical data. The data is formatted exactly like the RegularSeasonCompactResults data. Note that these games also include the play-in games (which always occurred on day 134/135) for those years that had play-in games. Thus each season you will see between 63 and 67 games listed, depending on how many play-in games there were.

Because of the consistent structure of the NCAA® tournament schedule, you can actually tell what round a game was, depending on the exact DayNum. Thus:

* DayNum=134 or 135 (Tue/Wed) - play-in games to get the tournament field down to the final 64 teams
* DayNum=136 or 137 (Thu/Fri) - Round 1, to bring the tournament field from 64 teams to 32 teams
* DayNum=138 or 139 (Sat/Sun) - Round 2, to bring the tournament field from 32 teams to 16 teams
* DayNum=143 or 144 (Thu/Fri) - Round 3, otherwise known as "Sweet Sixteen", to bring the tournament field from 16 teams to 8 teams
* DayNum=145 or 146 (Sat/Sun) - Round 4, otherwise known as "Elite Eight" or "regional finals", to bring the tournament field from 8 teams to 4 teams
* DayNum=152 (Sat) - Round 5, otherwise known as "Final Four" or "national semifinals", to bring the tournament field from 4 teams to 2 teams
* DayNum=154 (Mon) - Round 6, otherwise known as "national final" or "national championship", to bring the tournament field from 2 teams to 1 champion team

Special note: Each year, there are also going to be other games that happened after Selection Sunday, which are not part of the NCAA® Tournament. This includes tournaments like the postseason NIT, the CBI, the CIT, and the Vegas 16. Such games are not listed in the Regular Season or the NCAA® Tourney files; they can be found in the "Secondary Tourney" data files within Data Section 6. Although they would not be games you would ever be predicting directly for the NCAA® tournament, and they would not be games you would have data from at the time of predicting NCAA® tournament outcomes, you may nevertheless wish to make use of these games for model optimization, depending on your methodology. The more games that you can test your predictions against, the better your optimized model might eventually become, depending on how applicable all those games are. A similar argument might be advanced in favor of optimizing your predictions against conference tournament games, which might be viewed as fairly similar to NCAA® tournament games.

Data Section 1 file: SampleSubmissionStage1.csv

This file illustrates the submission file format for Stage 1. It is the simplest possible submission: a 50% winning percentage is predicted for each possible matchup.

A submission file lists every possible matchup between tournament teams for one or more years. During Stage 1, you are asked to make predictions for all possible matchups from the past four NCAA® tournaments (seasons 2014, 2015, 2016, and 2017). In Stage 2, you will be asked to make predictions for all possible matchups from the current NCAA® tournament (season 2018).

When there are 68 teams in the tournament, there are 68\*67/2=2,278 predictions to make for that year, so a Stage 1 submission file will have 2,278\*4=9,112 data rows.

* ID - this is a 14-character string of the format SSSS\_XXXX\_YYYY, where SSSS is the four digit season number, XXXX is the four-digit TeamID of the lower-ID team, and YYYY is the four-digit TeamID of the higher-ID team.
* Pred - this contains the predicted winning percentage for the first team identified in the ID field

Example #1: You want to make a prediction for Duke (TeamID=1181) against Arizona (TeamID=1112) in the 2012 tournament, with Duke given a 53% chance to win and Arizona given a 47% chance to win. In this case, Arizona has the lower numerical ID so they would be listed first, and the winning percentage would be expressed from Arizona's perspective (47%):

2012\_1112\_1181,0.47

Example #2: You want to make a prediction for Duke (TeamID=1181) against North Carolina (TeamID=1314) in the 2012 tournament, with Duke given a 51.6% chance to win and North Carolina given a 48.4% chance to win. In this case, Duke has the lower numerical ID so they would be listed first, and the winning percentage would be expressed from Duke's perspective (51.6%):

2012\_1181\_1314,0.516

Data Section 2 - Team Box Scores

This section provides game-by-game stats at a team level (free throws attempted, defensive rebounds, turnovers, etc.) for all regular season, conference tournament, and NCAA® tournament games since the 2002-03 season.

Team Box Scores are provided in "Detailed Results" files rather than "Compact Results" files. However, the two files are strongly related.

In a Detailed Results file, the first eight columns (Season, DayNum, WTeamID, WScore, LTeamID, LScore, WLoc, and NumOT) are exactly the same as a Compact Results file. However, in a Detailed Results file, there are many additional columns. The column names should be self-explanatory to basketball fans (as above, "W" or "L" refers to the winning or losing team):

* WFGM - field goals made (by the winning team)
* WFGA - field goals attempted (by the winning team)
* WFGM3 - three pointers made (by the winning team)
* WFGA3 - three pointers attempted (by the winning team)
* WFTM - free throws made (by the winning team)
* WFTA - free throws attempted (by the winning team)
* WOR - offensive rebounds (pulled by the winning team)
* WDR - defensive rebounds (pulled by the winning team)
* WAst - assists (by the winning team)
* WTO - turnovers committed (by the winning team)
* WStl - steals (accomplished by the winning team)
* WBlk - blocks (accomplished by the winning team)
* WPF - personal fouls committed (by the winning team)

(and then the same set of stats from the perspective of the losing team: LFGM is the number of field goals made by the losing team, and so on up to LPF).

Note: 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, of course. So if you want to know specifically about two-point field goals, you have to subtract one from the other (e.g., WFGM - WFGM3). And the total number of points scored is most simply expressed as 2\*FGM + FGM3 + FTM.

Data Section 2 file: RegularSeasonDetailedResults.csv

This file provides team-level box scores for many regular seasons of historical data, starting with the 2003 season. All games listed in the RegularSeasonCompactResults file since the 2003 season should exactly be present in the RegularSeasonDetailedResults file.

Data Section 2 file: NCAATourneyDetailedResults.csv

This file provides team-level box scores for many NCAA® tournaments, starting with the 2003 season. All games listed in the NCAATourneyCompactResults file since the 2003 season should exactly be present in the NCAATourneyDetailedResults file.

Data Section 3 - Geography

This section provides city locations of all regular season, conference tournament, and NCAA® tournament games since the 2009-10 season

Data Section 3 file: Cities.csv

This file provides a master list of cities that have been locations for games played.

* CityID - a four-digit ID number uniquely identifying a city.
* City - the text name of the city.
* 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.

Data Section 3 file: GameCities.csv

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, the NCAA® tourney, and other post-season tournaments, are all listed together.

* Season, DayNum, WTeamID, LTeamID - these four columns are sufficient to uniquely identify each game. Additional data, such as the score of the game and other stats, can be found in the corresponding Compact Results and/or Detailed Results file.
* CRType - this can be either Regular or NCAA® or Secondary. If it is Regular, you can find more about the game in the RegularSeasonCompactResults.csv and RegularSeasonDetailedResults.csv files. If it is NCAA®, you can find more about the game in the NCAATourneyCompactResults.csv and NCAATourneyDetailedResults.csv files. If it is Secondary, you can find more about the game in the SecondaryTourneyCompactResults file.
* CityID - the ID of the city where the game was played, as specified by the CityID column in the Cities.csv file.

Data Section 4 - Public Rankings

This section provides weekly team rankings for dozens of top rating systems - Pomeroy, Sagarin, RPI, ESPN, etc., since the 2002-2003 season

Data Section 4 file: MasseyOrdinals.zip containing MasseyOrdinals.csv

This zip file contains a large CSV file, listing out rankings (e.g. #1, #2, #3, ..., #N) of teams going back to the 2002-2003 season, under a large number of different ranking system methodologies. The information was gathered by Kenneth Massey and provided on his [College Basketball Ranking Composite page](https://www.masseyratings.com/cb/compare.htm).

Note that a rating system is more precise than a ranking system, because a rating system can provide insight about the strength gap between two teams. A ranking system will just tell you who is #1 or who is #2, but a rating system might tell you whether the gap between #1 and #2 is large or small. Nevertheless, it can be hard to compare two different rating systems that are expressed in different scales, so it can be very useful to express all the systems in terms of their ordinal ranking (1, 2, 3, ..., N) of teams.

* Season - this is the year of the associated entry in Seasons.csv (the year in which the final tournament occurs)
* RankingDayNum - this integer always ranges from 0 to 133, and is expressed in the same terms as a game's DayNum (where DayZero is found in the Seasons.csv file). The RankingDayNum is intended to tell you the first day that it is appropriate to use the rankings for predicting games. For example, if RankingDayNum is 110, then the rankings ought to be based upon game outcomes up through DayNum=109, and so you can use the rankings to make predictions of games on DayNum=110 or later. The final pre-tournament rankings each year have a RankingDayNum of 133, and can thus be used to make predictions of the games from the NCAA® tournament, which start on DayNum=134 (the Tuesday after Selection Sunday).
* SystemName - this is the (usually) 3-letter abbreviation for each distinct ranking system. These systems may evolve from year to year, but as a general rule they retain their meaning across the years. Near the top of the Massey composite page, you can find slightly longer labels describing each system, along with links to the underlying pages where the latest rankings are provided (and sometimes the calculation is described).
* TeamID - this is the ID of the team being ranked, as described in Teams.csv.
* OrdinalRank - this is the overall ranking of the team in the underlying system. Most systems provide a complete ranking from #1 through #351 (currently), but sometimes there are ties and sometimes only a smaller set of rankings is provided, as with the AP's top 25.

Disclaimer: you ought to be careful about your methodology when using or evaluating these ranking systems. They are presented on a weekly basis, and given a consistent date on the Massey Composite page that typically is a Sunday; that is how the ranking systems can be compared against each other on this page. However, these systems each follow their own timeline and some systems may be released on a Sunday and others on a Saturday or Monday or even Tuesday. You should remember that if a ranking is released on a Tuesday, and was calculated based on games played through Monday, it will make the system look unusually good at predicting if you use that system to forecast the very games played on Monday that already inform the rankings. To avoid this methodological trap, we have typically used a conservative RankingDayNum of Wednesday to represent the rankings that were released at approximately the end of the weekend, a few days before, even though those rankings are represented on the composite page as being on a Sunday. For some of the older years, a more precise timestamp was known for each ranking system that allowed a more precise assignment of a RankingDayNum. By convention, the final pre-tournament rankings are always expressed as RankingDayNum=133, even though sometimes the rankings for individual systems are not released until Tuesday (DayNum=134) or even Wednesday or Thursday. If you decide to use some rankings from these Massey Ordinals to inform your predictions, be forewarned that we have no control over when they are released, and not all systems may turn out to be available in time to make pre-tournament predictions by our submission deadline. In such a situation, you may wish to use the rankings from DayNum=128 or you may need to dig into the details of the actual source of the rankings, by following the respective links on the Massey Composite Page. We may also be able to provide partial releases of the final pre-tournament Massey Ordinals on the forums, so that as systems come in on Monday or Tuesday you can use them right away.

Data Section 5 - Play-by-play

This section provides play-by-play event logs for 99% of regular season, conference tournament, and NCAA® tournament games since the 2009-10 season - including plays by individual players.

Data Section 5 files: PlayByPlay\_201X.zip containing Events\_201X.csv and Players\_201X.csv

Each zip file (PlayByPlay\_2010.zip, PlayByPlay\_2011.zip, ..., PlayByPlay\_2017.zip) contains two CSV files, listing the play-by-play event logs for almost all games from that season. Each event is assigned to either a team or one of the team's players (by name). The players are listed by PlayerID within the Players csv file for that year, and the play-by-play events are listed (including a PlayerID) within the Events csv file for that year..

Data Section 5 file: Events\_201X.csv

* EventID - this is a unique ID for each logged event. The EventID's are different within each year, as are the PlayerID's. The events are sorted in approximate chronological order within each game, based on clock time, although when multiple events happen within the same clock time, the tiebreak for sorting is just the text EventType, so in some cases it may be impossible to determine the exact sequence of several tip-in attempts or free-throw-attempts that all happened at the same clock time.
* Season, DayNum, WTeamID, LTeamID - these four columns are sufficient to uniquely identify each game. The games are a mix of Regular Season, NCAA® Tourney, and Secondary Tourney games.
* WPoints, LPoints - whenever a scoring play happens (1 point, 2 points, or 3 points) the updated score is provided (from the perspective of the winning team (WPoints) and the losing team (LPoints), although of course during the game we didn't know yet that they were the winning team or losing team.
* ElapsedSeconds - this is the number of seconds that have elapsed from the start of the game until the event occurred. With a 20-minue half, that means that an ElapsedSeconds value from 0 to 1200 represents an event in the first half, a value from 1200 to 2400 represents and event in the second half, and a value above 2400 represents an event in overtime..
* EventTeamID - this is the ID of the team that the event is logged for, which will either be the WTeamID or the LTeamID.
* EventPlayerID - this is the ID of the player that the event is logged for, as described in the corresponding Players file.
* EventType - this is the type of the event that was logged (see listing below).

Event Types:

* assist - an assist was credited on a made shot
* block - a blocked shot was recorded
* steal - a steal was recorded
* turnover - a turnover was recorded
* timeout, timeout\_tv - a regular timeout or TV timeout was called
* foul\_pers, foul\_tech - a personal foul or technical foul was committed
* reb\_off, reb\_def, reb\_dead - an offensive rebound, defensive rebound, or dead-ball rebound was recorded
* sub\_in, sub\_out - a player entered or exited the game via a substitution
* made1\_free, miss1\_free - a one-point free throw was made or missed
* made2\_dunk, miss2\_dunk - a two-point field goal (dunk) was made or missed
* made2\_tip, miss2\_tip - a two-point field goal (tip-in) was made or missed
* made2\_lay, miss2\_lay - a two-point field goal (layup) was made or missed
* made2\_jump, miss2\_jump - a two-point field goal (jump shot) was made or missed
* made3\_jump, miss3\_jump - a three-point field goal (assumed to be a jump shot) was made or missed

Data Section 5 file: Players\_201X.csv

* PlayerID - this is a unique ID for each distinct player name. The PlayerID's are different within each year, as are the EventID's.
* Season - this is the year of the associated entry in Seasons.csv (the year in which the final tournament occurs)
* TeamID - this is the TeamID of the player's team, as described in Teams.csv.
* PlayerName - this is a text representation of the player's full name, in the format LAST\_FIRST, with underscores substituted in for spaces.

Note: there are errors within the events, in that they don't necessarily add up to the final stats for the game. Nevertheless, this was the highest quality data we could achieve for the near-complete set of games.

Data Section 6 - Supplements

This section contains additional supporting information, including coaches, conference affiliations, alternative team name spellings, bracket structure, game results for NIT and other postseason tournaments

Data Section 6 file: TeamCoaches.csv

This file indicates the head coach for each team in each season, including a start/finish range of DayNums to indicate a mid-season coaching change. For scenarios where a team had the same head coach the entire year, they will be listed with a DayNum range of 0 to 154 for that season. For head coaches whose term lasted many years, there will be many rows listed, most of which have a DayNum range of 0 to 154 for the corresponding year.

* Season - this is the year of the associated entry in Seasons.csv (the year in which the final tournament occurs)
* TeamID - this is the TeamID of the team that was coached, as described in Teams.csv.
* FirstDayNum, LastDayNum - this defines a continuous range of days within the season, during which the indicated coach was the head coach of the team. In most cases, a data row will either have FirstDayNum=0 (meaning they started the year as head coach) and/or LastDayNum=154 (meaning they ended the year as head coach), but in some cases there were multiple new coaches during a team's season, or a head coach who went on leave and then returned.
* CoachName - this is a text representation of the coach's full name, in the format first\_last, with underscores substituted in for spaces.

Data Section 6 file: Conferences.csv

This file indicates the Division I conferences that have existed over the years since 1985. Each conference is listed with an abbreviation and a longer name.

* ConfAbbrev - this is a short abbreviation for each conference; the abbreviation is used in some other files to indicate the parent conference of a team or of a conference tournament.
* Description - this is a longer text name for the conference.

Data Section 6 file: TeamConferences.csv

This file indicates the conference affiliations for each team during each season. Some conferences have changed their names from year to year, and/or changed which teams are part of the conference. This file tracks this information historically.

* Season - this is the year of the associated entry in Seasons.csv (the year in which the final tournament occurs)
* TeamID - this identifies the TeamID (as described in Teams.csv).
* ConfAbbrev - this identifies the conference (as described in Conferences.csv).

Data Section 6 file: ConferenceTourneyGames.csv

This file indicates which games were part of each year's post-season conference tournaments (all of which finished on Selection Sunday or earlier), starting from the 2001 season. Many of these conference tournament games are held on neutral sites, and many of the games are played by tournament-caliber teams just a few days before the NCAA® tournament. Thus these games could be considered as very similar to NCAA® tournament games, and (depending on your methodology) may be of use in optimizing your predictions. However, this is NOT a new listing of games; these games are already present within the RegularSeasonCompactResults and RegularSeasonDetailedResults files. So this file simply helps you to identify which of the "regular season" games since the 2001 season were actually conference tournament games, in case that is useful information.

* ConfAbbrev - this identifies the conference (as described in Conferences.csv) that the tournament was for.
* Season, DayNum, WTeamID, LTeamID - these four columns are sufficient to uniquely identify each game. Further details about the game, such as the final score and other stats, can be found in the associated data row of the RegularSeasonCompactResults and/or RegularSeasonDetailedResults files.

Data Section 6 file: SecondaryTourneyTeams.csv

This file identifies the teams that participated in post-season tournaments other than the NCAA® Tournament (such events would run in parallel with the NCAA® Tournament). These are teams that were not invited to the NCAA® Tournament and instead were invited to some other tournament, of which the NIT is the most prominent tournament, but there have also been the CBI, CIT, and Vegas 16 (V16) at various points in recent years. Depending on your methodology, you might find it useful to have these additional game results, above and beyond what is available from the NCAA® Tournament results. Many of these teams, especially in the NIT, were "bubble" teams of comparable strength to several NCAA® Tournament invitees, and so these games may be of use in model optimization for predicting NCAA® Tournament results.

* Season - this is the year of the associated entry in Seasons.csv (the year in which the post-season tournament was played)
* SecondaryTourney - this is the abbreviation of the tournament, either NIT, CBI, CIT, or V16 (which stands for Vegas 16).
* TeamID - this identifies the TeamID that participated in the tournament (as described in Teams.csv).

Data Section 6 file: SecondaryTourneyCompactResults.csv

This file indicates the final scores for the tournament games of "secondary" post-season tournaments: the NIT, CBI, CIT, and Vegas 16. The detailed results (team box scores) have not been assembled for these games. For the most part, this file is exactly like other Compact Results listings, although it also has a column for Secondary Tourney.

* SecondaryTourney - this is the abbreviation of the tournament, either NIT, CBI, CIT, or V16 (which stands for Vegas 16).

Data Section 6 file: TeamSpellings.csv

This file indicates alternative spellings of many team names. It is intended for use in associating external spellings against our own TeamID numbers, thereby helping to relate the external data properly with our datasets. Over the years we have identified various external spellings of different team names (as an example, for Ball State we have seen "ball st", and "ball st.", and "ball state", and "ball-st", and "ball-state"). Other teams have had more significant changes to their names over the years; for example, "Texas Pan-American" and "Texas-Rio Grande Valley" are actually the same school. The current list is obviously not exhaustive, and we encourage participants to identify additional mappings and upload extended versions of this file to the forums.

* TeamNameSpelling - this is the spelling of the team name. It is always expressed in all lowercase letters - e.g. "ball state" rather than "Ball State" - in order to emphasize that any comparisons should be case-insensitive when matching.
* TeamID - this identifies the TeamID for the team that has the alternative spelling (as described in Teams.csv).

Data Section 6 file: NCAATourneySlots

This file identifies the mechanism by which teams are paired against each other, depending upon their seeds, as the tournament proceeds through its rounds. It can be of use in identifying, for a given historical game, what round it occurred in, and what the seeds/slots were for the two teams (the meaning of "slots" is described below). Because of the existence of play-in games for particular seed numbers, the pairings have small differences from year to year. You may need to know these specifics if you are trying to represent/simulate the exact workings of the tournament bracket.

* Season - this is the year of the associated entry in Seasons.csv (the year in which the final tournament occurs)
* Slot - this uniquely identifies one of the tournament games. For play-in games, it is a three-character string identifying the seed fulfilled by the winning team, such as W16 or Z13. For regular tournament games, it is a four-character string, where the first two characters tell you which round the game is (R1, R2, R3, R4, R5, or R6) and the second two characters tell you the expected seed of the favored team. Thus the first row is R1W1, identifying the Round 1 game played in the W bracket, where the favored team is the 1 seed. As a further example, the R2W1 slot indicates the Round 2 game that would have the 1 seed from the W bracket, assuming that all favored teams have won up to that point. The slot names are different for the final two rounds, where R5WX identifies the national semifinal game between the winners of regions W and X, and R5YZ identifies the national semifinal game between the winners of regions Y and Z, and R6CH identifies the championship game. The "slot" value is used in other columns in order to represent the advancement and pairings of winners of previous games.
* StrongSeed - this indicates the expected stronger-seeded team that plays in this game. For Round 1 games, a team seed is identified in this column (as listed in the "Seed" column in the NCAATourneySeeds.csv file), whereas for subsequent games, a slot is identified in this column. In the first record of this file (slot R1W1), we see that seed W01 is the "StrongSeed", which during the 1985 tournament would have been Georgetown. Whereas for games from Round 2 or later, rather than a team seed, we will see a "slot" referenced in this column. So in the 33rd record of this file (slot R2W1), it tells us that the winners of slots R1W1 and R1W8 will face each other in Round 2. Of course, in the last few games of the tournament - the national semifinals and finals - it's not really meaningful to talk about a "strong seed" or "weak seed", since you would have #1 seeds favored to face each other, but those games are nevertheless represented in the same format for the sake of consistency.
* WeakSeed - this indicates the expected weaker-seeded team that plays in this game, assuming all favored teams have won so far. For Round 1 games, a team seed is identified in this column (as listed in the "Seed" column in the TourneySeeds.csv file), whereas for subsequent games, a slot is identified in this column.

Data Section 6 file: NCAATourneySeedRoundSlots.csv

This file helps to represent the bracket structure in any given year. No matter where the play-in seeds are located, we can always know, for a given tournament seed, exactly what bracket slot they would be playing in, on each possible game round, and what the possible DayNum values would be for that round. Thus, if we know when a historical game was played, and what the team's seed was, we can identify the slot for that game. This can be useful in representing or simulating the tournament bracket structure.

* Seed - this is the tournament seed of the team.
* GameRound - this is the round during the tournament that the game would occur in, where Round 0 (zero) is for the play-in games, Rounds 1/2 are for the first weekend, Rounds 3/4 are for the second weekend, and Rounds 5/6 are the national semifinals and finals.
* GameSlot - this is the game slot that the team would be playing in, during the given GameRound. The naming convention for slots is described above, in the definition of the NCAATourneySlots file.
* EarlyDayNum, LateDayNum - these fields describe the earliest possible, and latest possible, DayNums that the game might be played on.