**1. Data section 1:**

This section provides everything you need to build a simple prediction model and submit predictions. It's contain: 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 and Example submission file for stage 1.

**MTeams.csv and WTeams.csv:**

These files identify the different college teams present in the dataset.

* **TeamID** - a 4 digit id number, unique for each NCAA® men's or women's team. The men's ID's range from 1000-1999 and women's team ID's range from 3000-3999.
* **TeamName** - team name, 16 characters or fewer.
* **FirstD1Season** - the first season in dataset that the school was a Division-I school.
* **LastD1Season** - the last season in dataset that the school was a Division-I school. For any teams that are currently Division-I, they will be listed with LastD1Season=2025.

First abd Last D1Season column only appear in men's data.

**MSeasons.csv and WSeasons.csv:**

These files identify the different seasons included in the historical data, along with certain season-level properties. There are separate files for men's data (MSeasons) and women's data (WSeasons).

* **Season** - indicates the year in which the tournament was played.
* **DayZero** - date when season started, DayNum=0. So we can calucalate exact date of game, when can add to DayZero DaNum. 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** - each of the four regions in the final tournament is assigned a letter of W, X, Y, or 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 in a standardized way in other files, even if the region names change from year to year.

**MNCAATourneySeeds.csv and WNCAATourneySeeds.csv**

These files identify 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.

* **Season** - the year that the tournament was played in
* **Seed** - this is a 3-character or 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) tell 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. The "a" and "b" are assigned based on which Team ID is lower numerically.
* **TeamID** - this identifies the id number of the team

**MRegularSeasonCompactResults.csv and WRegularSeasonCompactResults.csv**

These files identify the game-by-game results for many seasons of historical data, starting with the 1985 season for men (the first year the NCAA® had a 64-team men's tournament) and the 1998 season for women. 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).

* **Season** - namely the year in which the final tournament occur.
* **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.
* **WTeamID** - this identifies the id number of the team that won the game, as listed in the "MTeams.csv" or "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.
* **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.
* **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 (or "away") team, this value will be "A". If it was played on a neutral court, then this value will be "N".
* **NumOT** - this indicates the number of overtime periods in the game, an integer 0 or higher.

**MNCAATourneyCompactResults.csv and WNCAATourneyCompactResults.csv**

These files identify the game-by-game NCAA® tournament results for all seasons of historical data. Because of the consistent structure of the NCAA® tournament schedule, you can generally tell what round a men's game was, just by looking at its day number. In general the men's schedule will be:

* 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.

**SampleSubmissionStage1.csv**

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

Submission file structure:

**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, the one represented above by XXXX.

**2. 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 2003 season (men) or since the 2010 season (women). 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.

* **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)

Files also include the same set of stats for losing team.

**3. Data Section 3 – Geography:**

This section provides city locations of all regular season, conference tournament, and NCAA® tournament games since the 2010 season.

**Cities.csv**

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

* **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.

**MGameCities.csv and WGameCities.csv**

These files identify 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 (known as secondary tournaments), are all listed together. File’s structure:

* **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 corresponding Regular Season Compact Results and Regular Season Detailed Results files. If it is **NCAA**, you can find more about the game in the corresponding NCAA Tourney Compact Results and NCAA Tourney Detailed Results files. If it is **Secondary**, you can find more about the game in the Secondary Tourney Compact Results file.
* **CityID** - the ID of the city where the game was played, as specified by the CityID column in the Cities.csv file.

4. **Data Section 4 - Public Rankings:**

This section provides weekly team rankings (men's teams only) for dozens of top rating systems - Pomeroy, Sagarin, RPI, ESPN, etc., since the 2003 season.

**MMasseyOrdinals.csv**

This file lists ordinal rankings (e.g. #1, #2, #3, ..., #N) of men's teams going back to the 2003 season, under a large number of different ranking system methodologies. File’s structure:

* **Season** - this is the year of the associated entry in MSeasons.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 MSeasons.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 generally 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.
* **TeamID** - this is the ID of the team being ranked, as described in MTeams.csv.
* **OrdinalRank** - this is the overall ranking of the team in the underlying system. Most systems from recent seasons provide a complete ranking from #1 through #351, but more recently they go higher because additional teams were added to Division I in recent years.

**5. Data Section 5 – Supplements:**

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

**MTeamCoaches.csv**

This file indicates the head coach for each team in each season, including a start/finish range of DayNum's to indicate a mid-season coaching change. For head coaches whose term lasted many seasons, there will be many rows listed, most of which have a DayNum range of 0 to 154 for the corresponding season.

* **Season** - this is the year of the associated entry in MSeasons.csv
* **TeamID** - this is the TeamID of the team that was coached, as described in MTeams.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.
* **CoachName** - this is a text representation of the coach's full name, in the format first\_last, with underscores substituted in for spaces, and all letters being lowercase.

**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.

**MTeamConferences.csv and WTeamConferences.csv**

These files indicate the conference affiliations for each team during each season. Some conferences have added or dropped teams from year to year, and these files track this information historically, for men's and women's teams separately.

* **Season** - this is the year of the associated entry in MSeasons.csv or WSeasons.csv (the year in which the final tournament occurs)
* **TeamID** - this identifies the TeamID (as described in MTeams.csv or WTeams.csv).
* **ConfAbbrev** - this identifies the conference (as described in Conferences.csv).

**MConferenceTourneyGames.csv and WConferenceTourneyGames.csv**

This file indicates which games were part of each year's post-season men's and women's conference tournaments (all of which finished on Selection Sunday or earlier), starting from the 2001 season for men's data and the 2002 season for women's data.

* **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 Regular Season Compact Results and/or Regular Season Detailed Results files.

**MSecondaryTourneyTeams.csv and WSecondaryTourneyTeams**

This file identifies the teams that participated in post-season men's or women's tournaments other than the NCAA® Tournament (such events would run in parallel with the NCAA® Tournament).

* **Season** - this is the year of the associated entry in the Seasons file (the year in which the post-season tournament was played)
* **SecondaryTourney** - this is the abbreviation of the tournament, such as NIT or WNIT.
* **TeamID** - this identifies the TeamID that participated in the tournament (as described in MTeams.csv or WTeams.csv).

**MSecondaryTourneyCompactResults.csv and WSecondaryTourneyCompactResults**

This file indicates the final scores for the tournament games of "secondary" post-season tournaments. For the most part, this file is exactly like other Compact Results listings, although it also has a column for Secondary Tourney. Also note that because these games are played after DayNum=132, they are NOT listed in the Regular Season Compact Results file.

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

**MTeamSpellings.csv and WTeamSpellings.csv**

These files indicate alternative spellings of many team names. They are intended for use in associating external spellings against our own TeamID numbers, thereby helping to relate the external data properly with our datasets.

* **TeamNameSpelling** - this is the spelling of the team name. It is always expressed in all lowercase letters.
* **TeamID** - this identifies the TeamID for the team that has the alternative spelling (as described in MTeams.csv or WTeams.csv).

**MNCAATourneySlots and WNCAATourneySlots**

These files identify the mechanism by which teams are paired against each other, depending upon their seeds, as the tournament proceeds through its rounds.

* **Season** - this is the year of the associated entry in MSeasons.csv or WSeasons.csv
* **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.
* **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 MNCAATourneySeeds.csv or WNCAATourneySeeds.csv file), whereas for subsequent games, a slot is identified 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.
* **WeakSeed** - this indicates the expected weaker-seeded team that plays in this game, assuming all favored teams have won so far.

**MNCAATourneySeedRoundSlots.csv**

This file helps to represent the men's 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.

* **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.
* **EarlyDayNum, LateDayNum** - these fields describe the earliest possible, and latest possible, DayNums that the game might be played on. Many tournament rounds span two days' time and so the expected day number for a game to be played on would be one of those two days.