After discovering a spreadsheet with a lot of data surrounding the past NBA season results, I noticed many changes that would be necessary to use this data for analysis.

The steps I took and reasons behind these steps are below:

1. The date column was written as just integers and was missing the “0” before any dates for months like January, February, etc (01, 02…) Padding the “0” into this column involved locating any values from the date column that only had three values and then adding the “0” to the beginning.
2. Each game and the results of the games were separated in groups of two rows. Therefore, I needed to separate each teams individual results and keep the correct matchups. To do this, I created two separate tables based upon home teams and visitor teams. After separation, the next step was merging the data tables so that looking across one row one would read the following: the date of the game, teams that played, the projected lines and the actual results of the game.
3. After merging the tables together and having one large dataset to work with, I discovered that the spread for each game was assigned to whichever team was the favorite. This made for a mix up of spread data and over/under data. I then swapped the values to get all spread data in one column and all over/under data in one column.
4. One of the final steps from a numerical standpoint was getting the spread value properly aligned to the favorite. I used the favorite’s Money Line as an index to make the spread values negative for the favorites, the way that Vegas Books provides lines.
5. Renaming the columns for readability sake was the next step that was necessary for the dataset.
6. The final steps involved removing columns that were not necessary for my future analysis. This way I have a clean, ready to use dataset for the remainder of my project.

The cleanup of my dataset was very tedious and time consuming to get the information in the correct formatting for analysis.