Data Wrangling report

1/20/19

Problem

Which regular-season team statistics correlate most with wins in the NCAA Tournament?

Data

- Regular seasons stats for all 351 men's college basketball team from 2014-2018 (source: SportsReference)
- Post-season results for the 68 teams that made the NCAA Tournament in each of those years (source: Data.world user-submitted table, cross referenced with SportsReference)

Goal

Using data wrangling techniques learned in Unit 3, my goal was to find data on how teams performed in various categories through the regular season (field goal percentage, assists per game, etc.) and how teams that made the NCAA Tournament performed in the "Big Dance." Then, I set out to clean and merge those data sets to position them for data analysis -- seeing which stats correlate the most with wins.

Major steps

Finding and cleaning the data took about two dozen hours. There were parts that went smooth and parts that stumped me (see *'s below). There were some "eureka!" moments and moments where I had to close my laptop lid for that night. There were considerations that arose later in the process that prompted me to go back and restructure chunks of my code. Ultimately, with some help from my mentor and online forums, I was able to get clean and join the two data sets into the format that I wanted. Here's the gist of what I did, along with a few notes:

- **Find and download stats** - The SportsReference had four categories for its downloadable team season stats data - team basic, team advanced, opponent basic and opponent advanced. I had to download all 20 and convert to .xlsx format. Its tournament results data was not in a downloadable format, so I copied and pasted the results into an Excel workbook and did some light cleaning there. I later found a tourney results data set on Data.world that would work, which I compared with SportsReference results.
- **Clean/merge stat data** - I primarily worked with 2018 data. Since there were four categories for each I wanted to merge them all. I had to add columns, remove redundant columns and rename offensive and defensive stat columns that had the same name.
- **Clean tourney data** - This data proved a little more difficult to handle that the season stat data. I had to display, with 1s and 0s, who the winners were in all 67 games. Then I needed to make binary columns for each round (Sweet 16, Elite 8, etc.) to show what each team's record was in that round(*). I finally added a column to show how many total wins in the tourney for each team(*).
- **Merge stats and tourney data** - Biggest issue here was that some school names for Data.world were different from some school names for SportsReference. I had to find them use SportsReference as the default(*).