Matt Fannin

College Basketball Data Analysis

Summer Project 1: College Basketball Using Pandas

Research Questions/Goals:

Study University of Michigan's data in more depth

Study the BIG 10 conference, find out which schools are the most successful

Find out which of the power 5 conferences is the most successful, wins the most

Offense vs Defense, which will lead to winning more games

Data Sources:

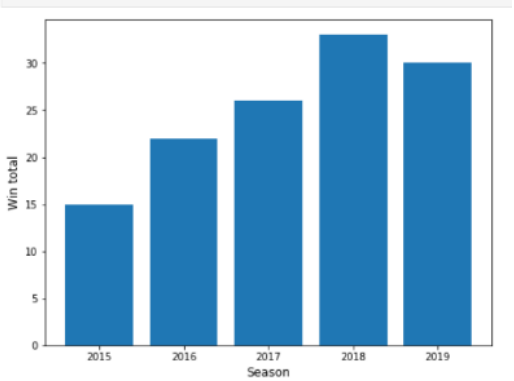
I got the data from Kaggle, and was a csv format. It contained 24 columns that housed various statistics about a team's season. The data spanned from 2015-2019, so five seasons worth of data was captured per school.

Data Manipulation:

The dataset was already very clean, when it was downloaded. To begin I first read in the csv into JupyterLab, and created a data frame using pandas. I then set the index to be the team. This was all the set up I did, before beginning to analyze.

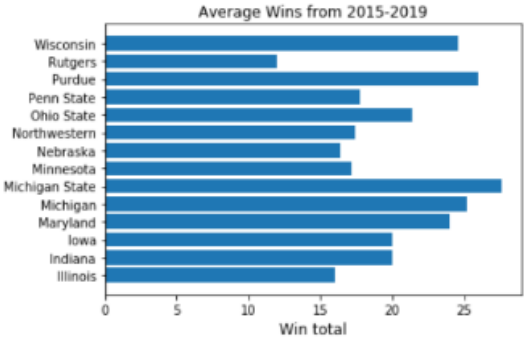
Analysis and Visualizations:

To begin, I wanted to start narrow and then study broader topics. So, I begun by creating a new data that contained only the University of Michigan's data. I wanted to first find out which of the five seasons was their best (based upon win total). The bar plot clearly shows that season was in 2018, where they accumulated over 30 wins. Next, I wanted to try and find out why the team was so successful that season. I did this by taking all the statistics that were given in the dataset ex: (field goal

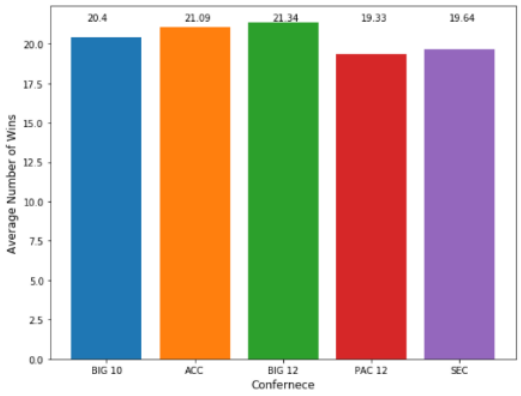


percentage, offensive rebound percentage, tempo, and others) and ranking them using pandas rank method. A five represents the highest value, and a one represents the lowest value in each column. I made a new data frame that held those ranks, and made a new column that had the average rank. I found in 2018, that rank was 3.1. This was the third highest behind 2017 and 2016. After looking at the data more, it led me to believe the team in 2018 was successful because they were so well-rounded and didn't have a one in any of the statistics.

I then shifted my attention to the BIG 10 conference as a whole. I filtered through the dataset by checking the "CONF" column for "B10", then created a data frame with each team's average statistics over the last five seasons, and using groupby method on the "TEAM" column to create an easy to look at dataset. Next, I wanted to plot the average number of wins over the last five years, and find the top five teams with the most wins. I sorted the data frame in descending order based on the "W" column, grabbed the top 5 and saved those to a new variable. I found that Michigan State average the most wins per season at 27.6, followed by: Purdue, Michigan Wisconsin, and Maryland. Following a similar strategy as stated above I ranked each school in the statistics in the dataset, to see if Michigan State had the highest average. I found that they did not, Purdue had the highest average at 3.6, followed by Michigan State at 3.13. However, after looking at the rankings I saw that Michigan State shot the highest field goal percentage, and allowed the lowest field goal percentage. These are key statistics that could be the reason why they win so many games and lead the BIG 10 in wins per season, on average.



Now I wanted to see which power 5 conference on average won the most games during the five-year span. To accomplish this, I followed the same process of collecting the BIG 10 data. I filtered through the "CONF" column, then used .mean() to average the five years of statistics together. I saved those five values to variables I used Matplotlib to create a bar plot. Although all the conferences are similar in average win total, the BIG 12 conference teams average 21.34 wins. On the other hand, PAC 12 teams average the fewest wins per season at 19.33.



Lastly, I wanted to dive into how offense and defense affect the chances of winning. I started to do this by creating two new data frames, one containing offense (wins, 2-point shooting percent, 3-point shooting percent, field goal percent, and adjusted offensive efficiency) the other containing defense (wins, 2-point shooting percent allowed, 3-point shooting percent allowed, field goal percent allowed, and adjusted defensive efficiency). I created a data frame of correlation values using .corr() in pandas. On offense, it appears the higher the percentage of made 2 and 3-pointers leads to a higher correlation of a higher total field goal percentage. A higher field goal percentage correlates to more efficient offense, and that has a .75 value of correlating to a win. On the other side of the ball, it shows that if a team defends against 2-pointers strongly, it makes their defense stronger. A higher defensive efficiency is negatively correlated to winning. It has a value of -.69, meaning a team is more likely to win when they hold the opponent to less points. It appears that offensive efficiency has a stronger correlation to winning a basketball game, based upon my findings from the data.

Skills Practiced/ Things Learned:

- Practiced working with pandas
- Reading documentation to find out how to work with methods in pandas and Matplotlib
- .corr()
- .rank()