**Statistical/Hypothetical Question**

* What factors contribute to college football teams winning more than others?

**Outcome of EDA**

During the EDA, I found a lot of information that informed my final analysis. The histogram of every metric showed me the distribution of the data, as well as any potential outliers. There were five teams that were extreme outliers for points scored. However, I decided to leave them in the analysis since those teams also had high win totals, so may lead to incorrect conclusions if they were excluded. I also built a descriptive statistics table, which showed the histogram data in more details and showed the spread of the data for each variable.

I also completed a PMF to see if the most powerful conference had a difference distribution than other conferences, which I didn’t find to be the case for the most part. I also conducted a CDF to look at the cumulative spread of the data for points scored and points given up, which showed, like my histograms found as well, that there is a bigger variability between points scored by teams versus points given up. I also looked at correlations between metrics to see relationships. While correlations don’t necessarily mean causation, it was important to see how variables were related to each other to see what impact they might have on my final analysis. I saw that recruiting was highly correlated with wins, and a smaller relationship between teams points scored versus points given up.

After doing my exploratory data analysis, I completed a multiple regression analysis and found that the overall relationship was significant, with the biggest impact was that points scored impacted win totals. While this is not groundbreaking, it was good to get confirmation that my model was able to detect this relationship. I also found in my model looking at defense only metrics, that recruiting was significant as well.

**What do you feel was missed during the analysis?**

While I learned a lot during the analysis, there were several aspects I missed during the review. While this analysis was purposely done at a high level, I feel like breaking down the data further would have led to more significant findings. There were also some correlations I could have done and more hypothesis testing if I had more detailed data, such as seeing which positions on defense for recruiting had the biggest impact on win total.

**Where there any variables you felt could have helped in the analysis?**

While I feel like I had the variables I needed for the analysis, I believe it would have been helpful if I had broken out some of the variables some more. For example, I only looked at major conference football teams, but I included the non-conference games in the win total, which may have skewed the data for teams that played easier schedules. Similarly, when looking at the recruiting data, I didn’t break out the recruiting ranks by offense vs defense. This would have been helpful when looking at how much impact offense or defense had in winning games.

**Where there any assumptions made you felt were incorrect?**

I believed when selecting the metrics that success rate would have a bigger impact on the overall model than it did. Part of an analysis is making assumptions and letting the data prove you wrong or right, so I don’t think it was a mistake to select these variables, but I learned they weren’t as impactful as I thought they would be at the beginning of the analysis. So if there was more time, I may have gone back and tried to select other metrics to include that would have been more helpful predicting wins.

**What challenges did you face? What did you not fully understand?**

I don’t believe I had a lot of challenges, since I luckily selected a data set that was clean for the most part, with no missing or partially missing data for any of the teams. As I mentioned above, there were some changes I would have made in the data if I knew then what I knew now, like filtering the data down to only conference performance, to remove any influence weaker non-conference opponents had on the metrics. It also helped that I picked a subject (college football) that I have a lot of knowledge about, whereas if I picked a topic that was new to me, I probably would have had to spend a lot more time on the metric definitions and data understanding to make sure I was measuring the data correctly.

Going into the project, I didn’t fully understand how much work would go into the EDA. I did several iterations of metrics to look at relationships, and I know there is a lot more drill-down and comparisons I could do if there was more time in the quarter. There are still a lot of concepts covered in the course I didn’t get a chance to practice on the data, like time series analysis, but this project has helped me understand what goes into a data science project and will help shape my planning and EDA process going into future projects, both in the classroom and in the real world.