Analysis of College Basketball Seasons

Team Member: Winston Vu

Introduction and Data Source

The game of basketball has evolved throughout the years, and there are changes within the style of play. In the past, basketball has usually been a slow paced game revolving around ball movement, passing the ball to the post, and 2-point shooting. Nowadays, the game is fast paced, revolving around fast ball movement and 3-point shooting. This style of play has changed within the National Basketball League (NBA) and college basketball. Notably, this style of game has changed by Golden State Warriors guard Stephen Curry when his 3-point shooting prowess took over the game, and now, many coaches strive for this style of play as well as the players.

The dataset I have chosen is a college basketball dataset from Kaggle, which includes all stats to show how each team played per season and how each team played during that season. The columns within the dataset are: team, conference, games played, wins, adjusted offensive and defensive efficiency, power rating, effective field goal percentage for offense and defense, turnover rate, steal rate, offensive and defensive rebound rate, free throw rate offensively and defensively, 2-point percentage offensively and defensively, 3-point percentage offensively and defensively, adjusted tempo, wins above bubble, March Madness seed, and year of season. Originally, I wanted to do the years 2017 to 2021, but since COVID-19 was a barrier for college basketball in the year 2020, I decided to use the years 2016 to 2019 because there might have been teams that were affected by COVID-19, the data would have been skewed or varied, and Stephen Curry won the most valuable player (MVP) in 2015 season.

Since the style of play has changed, I wanted to explore how the evolution of the game has changed within the world of college basketball. College players have to adapt to the new style of play in order to fulfill their dreams of possibly playing basketball at a higher level. Also, teams have to adjust to the new style of play, and it may affect how coaches draw up a game plan from their original style of play. The better the game plan means the better the team plays and the better the teams are long-term. Business-wise, teams are implanting a new game plan to win the March Madness championship, get highly rated recruits, and have a reputation as an outstanding basketball school. The better a team's success is means the better the likelihood the program evolves into a wealthy program.

Analysis Questions

There are multiple questions I wanted to uncover with this dataset:

- Win Percentage
    - Which team had the highest win percentage within the 4 years?
    - Out of the March Madness Champions within the 4 years, which team had the highest average win percentage?

        - It is of note that the teams that won within the 4 years are:
            - Villanova (2016 and 2018)
            - North Carolina (2017)
            - Virginia (2019)
    - What was the average win percentage of the University of Texas at Austin? Syracuse University?
        - Used for personal fun since I go to these schools
- 3-point Shooting
    - Which teams had the highest 3-point percentage?
    - Using a multiple regression analysis, is 3-point percentage (base on offense and defense) and significant factor in win percentage?
- Stats
    - What's the correlation between the stats given and win percentage?

Method of Analysis

The method of analysis used will be data importation, data cleansing, and data preparation. The packages that will be use for the analysis are Pandas, NumPy, matplotlib, seaborn, Scikit-learn, and statsmodel.api.

Data importation

To import the data, I used Pandas and its function "read_csv" to import the data into Jupyter Notebook.

```
(1406, 24)
```

Using the 'shape' function, the dataset has 1406 rows with 24 columns totaling to 33,744 points of data.

Data Cleansing

For the data cleaning, there was note of NAs within the data. To remove these NAs, I switched the NAs to 0 using the 'fillna(0)' function. Also, since I wanted to only use the years 2016 to 2019, I had to remove the rows. First, I reorganized the data by year using the 'sort_values' function and setting the ascending portion to 'False' in order to have the years start from 2019 and downward. Then, I removed the rows from 2013 to 2015 by choosing the column and using '!=' on the year of choice for removal. Lastly, I had to delete the columns I did not think would have an impact on the final outcome. The columns deleted were ' BARTHAG', 'WAB', and 'SEED'.

Data Preparation

For data preparation, I renamed the columns in order to easily understand the data as well as easily accessing the data. In order to rename the columns, I used 'rename(columns={})'. The column names are as followed:

- TEAM -> Team
- CONF -> Conf
- G -> Games_Played
- W -> Games_Won
- ADJOE -> Adj_Off_Eff
- ADJDE -> Adj_Def_Eff
- EFG_O -> Eff_FG%_Off
- EFG_D -> Eff_FG%_Def
- TOR -> Turnover_Rate
- TORD -> Steal_Rate
- ORB -> Off_Reb_Rate
- DRB -> Def_Reb_Rate
- FTR -> FT_Rate_Off
- FTD -> FT_Rate_Def
- 2P_O -> 2_Point%_Off
- 2P_D -> 2_Point%_Allowed
- 3P_O -> 3_Point%_Off
- 3P_D -> 3_Point%_Allowed
- Adj_T -> Adj_Tempo
- YEAR -> Year

After renaming the columns, I added a column with the name 'Win%'. In order to do this, I created a column within the original data frame, and I divided the 'Games_Won' column by the 'Games_Played' column.

```
Post preparation shape:
 (1406, 22)
```

After adding this column, the rows stayed the same at 1406, but the columns decreased to 22 due to removing 3 columns and adding 1 more. The total amount of data points for the final data frame is 30,932.

Question 1: What team had the highest win percentage within the 4 years? The lowest?

```
The team with the highest win percentage was Gonzaga at 94.87% in the year 2017.
The team with the lowest win percentage was Chicago St. at 3.45% in the year 2016.
```

Based on the results, it was not surprising to see Gonzaga to have the highest win percentage, as they play in an easy conference, have great coaching (Mark Few has been a coach there for a decade), and have been historically great. For the worst win percentage, it did not surprise me that Chicago St. was the answer because it did not occur to me that Chicago St. was an actual school.

Question 2: Since shooting 3-pointers is the new style play of basketball, which teams had the highest 3-point percentage within the 4 years? Which teams are bad at defending the 3?

```
Highest 3 point percentage teams:
                Team  3_Point%_Off
2181    Michigan St.          43.4
2205       Marquette          42.9
1109          Lehigh          42.4
521   William & Mary          42.2
45           Oklahoma          42.2


The worst 3 point defending teams:
                Team  3_Point%_Allowed
258     USC Upstate              43.1
364   Southern Utah              42.2
1451     Pepperdine              42.0
1343         Denver              41.8
551            Rice              41.8
```

Comparing these results, the best 3-point shooting teams have about the same percentages with the worst 3-point defending teams. Looking at a coach's perspective, this would be useful for game planning against an opponent. If a coach knows a team is good at shooting 3-point shots, then they should practice defending the 3, and if a team is bad at defending the 3-point shot, then the coach should make the players practice shooting the 3. It is of note that only 2 power-5 teams are noted as teams that shot the 3-pointer the best (Michigan St. and Oklahoma). This shows that the game of college basketball has either teams that shoot the 3-pointer a lot and very well or teams that shoot the 3-pointer a little but very well.

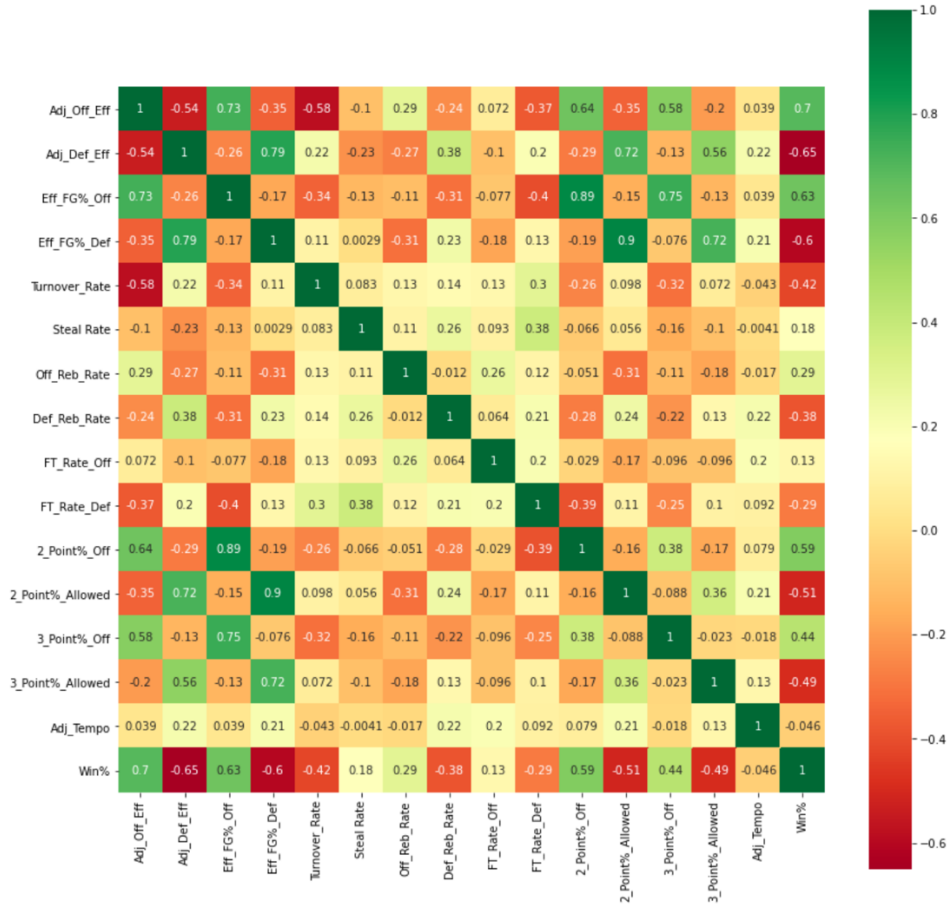
Question 3: What is the change in win percentage for my alma mater, Texas, and my current university, Syracuse? How has the win percentage change each season for each of the champions (Villanova, Virginia, and Baylor)?

```
The average win percentage for Texas is 51.64% from the year 2016 to 2019.
The average win percentage for Syracuse is 59.76% from the year 2016 to 2019.
The average win percentage for UNC is 79.49% from the year 2016 to 2019.
The average win percentage for Villanova is 84.65% from the year 2016 to 2019
The average win percentage for Virginia is 82.33% from the year 2016 to 2019.
```

Based on the results, Villanova has the highest win percentage. This does not come as a surprise as they have won 2 March Madness Championships. It is of note that the winners of March Madness had very high win percentages. If there was a scatterplot amongst all the winners of March Madness and their win percentage, then they would likely all be in one spot. Another thing of note is that the win percentage for Texas and Syracuse are very low to those that have won the tournament. Both teams are known as blue-bloods, which entails they have a winning history of the sport and/or lots of money for the sport. Despite their reputation and advantages, they have struggled with winning games, which is also noticeable during football season as well.

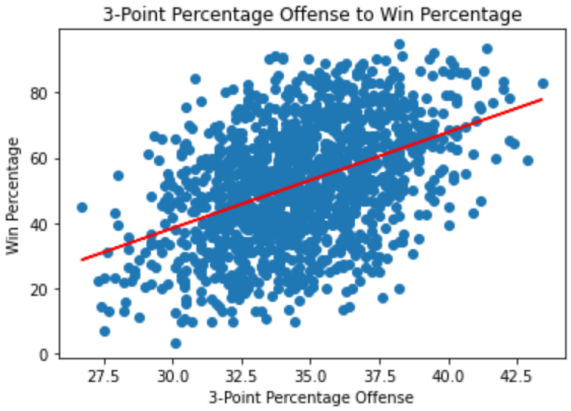
Question 4: What stats are mainly correlated to a higher win percentage?

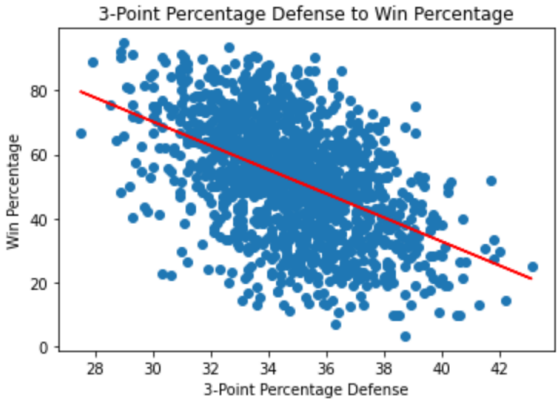
```
Win Percentage Correlation:
 Win%                1.000000
Adj_Off_Eff          0.695083
Eff_FG%_Off          0.634095
2_Point%_Off         0.593644
3_Point%_Off         0.442494
Off_Reb_Rate         0.291860
Steal Rate           0.175753
FT_Rate_Off          0.129505
Adj_Tempo           -0.045864
FT_Rate_Def         -0.285449
Def_Reb_Rate        -0.382775
Turnover_Rate       -0.420977
3_Point%_Allowed    -0.492452
2_Point%_Allowed    -0.511597
Eff_FG%_Def         -0.602441
Adj_Def_Eff         -0.651037
Name: Win%, dtype: float64
```

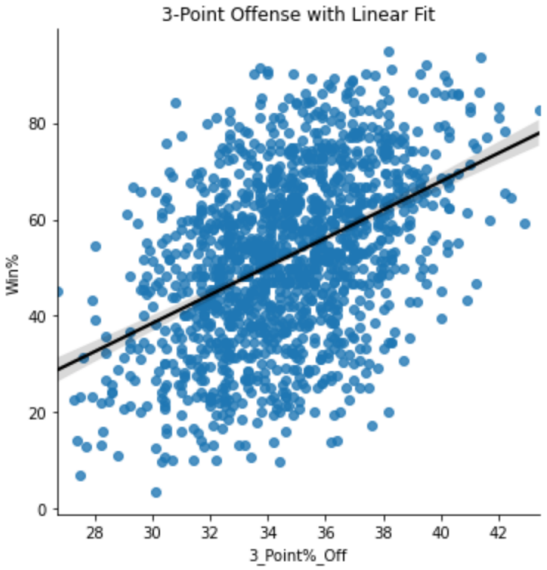


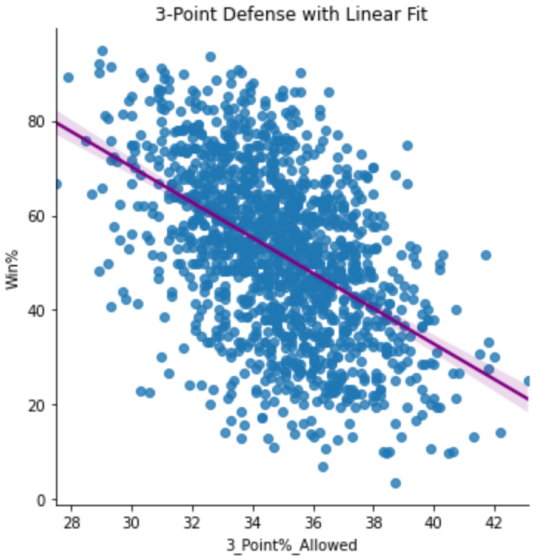
It does not come as a surprise that the adjusted offensive and defensive efficiencies would have the highest correlations. A notable result is that the stats on 2-point shots were correlated more with win percentage than any stats on the 3-point shots. This may be due to the idea that teams still shoot 2-point shots more efficient than 3-point shots. It is possible that 3-point percentage might have a higher correlation when compared to total points instead of win percentage. An interesting idea that can be drawn from the correlation results is that not all offensive and defensive stats equal amongst each other. For example, on the free throw stats, the offensive correlation to win percentage was 0.13 and the defensive, or amount allowed, correlation to win percentage was -0.29. It shows that some offensive and defensive stats that correspond with each other might be more important to win percentage. Free throws are important to the score of basketball, but when a team is trying to win, the team should avoid fouling players in order to avoid free throws.

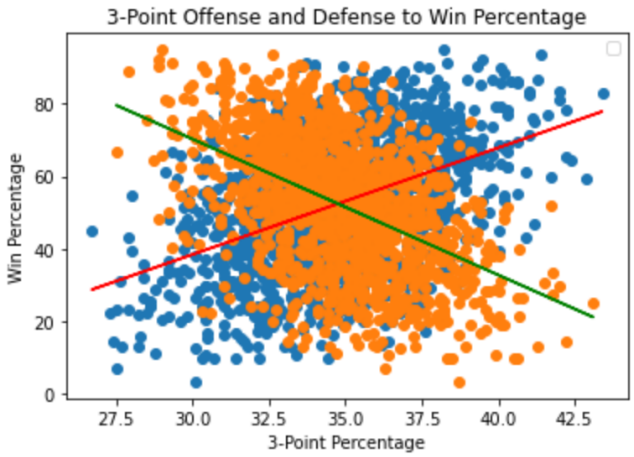
Question 5: Performing a multiple linear regression on 3-point percentage offensively and defensively to win percentage, which ones are statistically significant?











```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Win%   R-squared:                       0.429
Model:                            OLS   Adj. R-squared:                  0.428
Method:                 Least Squares   F-statistic:                     526.0
Date:                Mon, 06 Dec 2021   Prob (F-statistic):          3.40e-171
Time:                        21:49:18   Log-Likelihood:                -5644.4
No. Observations:                1406   AIC:                         1.129e+04
Df Residuals:                    1403   BIC:                         1.131e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              80.6017      7.171     11.240      0.000      66.535      94.669
3_Point%_Allowed   -3.6673      0.153    -23.903      0.000      -3.968      -3.366
3_Point%_Off        2.8652      0.134     21.370      0.000       2.602       3.128
==============================================================================
Omnibus:                        3.056   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.217   Jarque-Bera (JB):                3.121
Skew:                          -0.108   Prob(JB):                        0.210
Kurtosis:                       2.919   Cond. No.                         986.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The linear regression results show an R-squared value of 0.429, meaning 42.9% of the data can fit the regression model. The p-values associated with each variable is 0.00, meaning if the alpha value is 0.05, then these values are not independent and have statistical significance. It is interesting that all these p-values are 0.00 because it is particularly impossible to have that kind of value.

```
p-value:
 const                3.963112e-28
3_Point%_Allowed    3.278784e-106
3_Point%_Off         6.031372e-88
```

After using the 'pvalues' function in order to find each variable's p-value. It is shown that these p-values are actually very lowm and they are still statistically significant towards the model. 3-pointers, offensively and defensively, have an impact towards win percentage, which is no surprise. Although the values are statistically significant, the model is still bad. Since the R-squared value is below 0.70, which is the general consensus for a bad R-squared value, there is not much explanation from the independent variable to the dependent variable.

```
              Feature         VIF
0               const  401.528824
1   3_Point%_Allowed    1.000528
2      3_Point%_Off     1.000528
```

A strange thing that occurred when I played around with variables was an issue of multicollinearity. After various combinations, the 3-pointer offense and defense was the only combination that did not have an error message issued for multicollinearity. To find whether or not variables have a high level of multicollinearity is variance inflation factor (VIF). The VIF above is for 3-pointers only, and the variance inflation factors for the 3-point variables is below 10, so there was no issue of multicollinearity. The VIF for the constant, or intercept, is substantially high, but it is not important for the multicollinearity between variables.

Conclusions

The dataset provides valuable data for coaches, businesses, or ordinary spectators. The dataset can provide trends for each college basketball season in order for one to understand what occurred during the season and see if there was any changes of note that can be significant for the next season. These can be used during in game situations, for coaching evaluations, create a March Madness prediction simulator, or for other various entities.

In my search for trends, I wanted to see if the 3-point game had an overall effect on a team's win percentage. Although the idea sounded plausible, it did not have enough information to consider the model useful. Adding information about the amount of 3-pointers shot and 3-pointers made can be useful in determining a team's 3-point percentage to have an impact. Although the results of the correlation can give us an idea of the impact of certain variables, there is not enough information on most variables to create a concrete solution.

This dataset can provide results and information we desire to look at. On a larger scale, it can be flawed because some information is not fully in context. There can be further investigation on the data, but supplemental information can be useful for more complex designs and more complicated analysis.

Further Analysis/Next Steps

An idea that sounds interesting for a future analysis would be comparing stats, win percentage, or both to Name, Image, and Likeness (NIL). The NIL process is relatively new, and it is a system, or process, that helps collegiate athletes earn money in college. It would be interesting to see how a player's stats, team's stats, or win percentage would affect the NIL deal a player receives. Another entity that can be created is creating a linear regression model that helps predict the NIL deal a player receives based on stats. The NIL process just recently started in 2021, so it may take time to receive or being able to receive the data, gather sufficient amounts of data, and confirm that the data is correct.

Another idea that can be fun to analyze would be a team's play and high school recruiting. When a high school player gets recruited, they are denoted by the amount of 'stars' they have. For example, a 5 star player, which is the max, would want to go to school's with high pedigree

and/or success in their program. It would be interesting to gather a recruiting database and see whether or not a team's success would match a high-profile recruit's standards.

Lastly, it would be interesting to create a March Madness simulator to see who wins based off stats. Since March Madness will be continuing again for the 2021-2022 season, it would be interesting to see which statistical models and algorithms would create the perfect model. A possibility is looking at past March Madness results and models to see how each team won and how the distribution of teams goes.

References

College basketball dataset
https://www.kaggle.com/andrewsundberg/college-basketball-dataset?taskId=661

Correlation matrix heatmap
https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

Linear regression OLS table/results
https://datatofish.com/statsmodels-linear-regression/

Multicollinearity and Variance Inflation Factor
https://towardsdatascience.com/everything-you-need-to-know-about-multicollinearity-2f21f082d6dc

Linear fit plot
https://seaborn.pydata.org/generated/seaborn.regplot.html