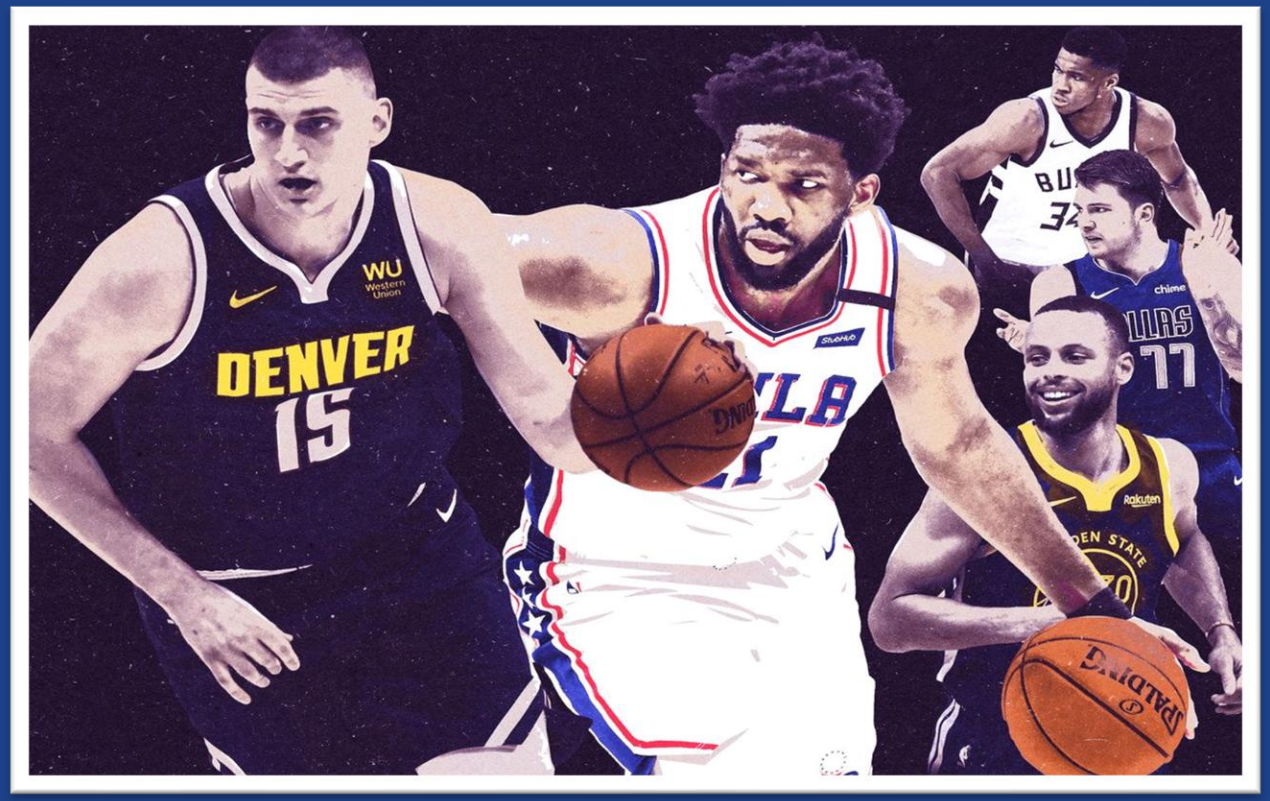
Path to the NBA:

A detailed look into ESPN High School Basketball Player ratings predictability on making it to the NBA



Prepared by:

Diane Onguetou

Tanner Olson

Date:

May 24th, 2021

1. Motivations

Nature of the Project

The odds of one day playing professional basketball in the NBA are low. It takes grit, determination, often times some good genetics, and maybe even some luck thrown into the mix to make it to the top. It has been noted that high school seniors who play basketball for their school's varsity team have a mere 0.03% (Norwichcsd) chance of one day getting drafted into the league.

With over 500k boys participating in high school basketball every year in the United States, picking out future NBA players is tough. In order to highlight top prospects , Entertainment and Sports Programming Network (ESPN), ranks what they believe are the top 100 high school prospects each year.

As ESPN's ranking algorithm is proprietary, this report will "audit" its model to gauge its effectiveness in predicting future NBA players.

With only 60 players getting drafted into the NBA each year, we will merge ESPN ranking data with NBA draft data to begin to evaluate ESPN's ranking algorithm's effectiveness. We will also investigate player characteristics to highlight any potential patterns in high school players that eventually make it into the NBA.

Specific Questions and Goals

As fans of sports, we are interested in potential tools that help signal future player performance. As ESPN's player rankings are constantly regarded as the source of truth for gauging skill amongst peer groups, we will be looking into the validity of this notion.

Through this analysis, we hope to be able to determine

i. Specific characteristics that top players have in common that make it to the NBA, and how many years it takes players to get drafted

ii. Overall, how effective is ESPN's high school basketball rankings at predicting whether a player eventually gets drafted into the league.

Possible Applications

Depending on the results, this analysis could act as a spring board for additional analyses around what players should focus on in order to get ranked and what college/NBA teams should focus on in order to narrow in on top players.

Recruiting Database
2021 ESPN 100

RK	PLAYER	POS	HOMETOWN	HT	WT	STARS	GRADE	SCHOOL
1	Chet Holmgren	C	Minneapolis, MN Minnehaha Academy	7'1"	195	★★★★★	97	GONZAGA SIGNED
2	Jaden Hardy	SG	Henderson, NV Coronado High School	6'4"	185	★★★★★	97	List
3	Paolo Banchero	PF	Seattle, WA O'Dea High School	6'9"	225	★★★★★	97	DUKE SIGNED
4	Patrick Baldwin Jr.	PF	Sussex, WI Hamilton High School	6'9"	220	★★★★★	97	MILWAUKEE COMMITTED 05/12/2021
5	Jabari Smith	PF	Tyrone, GA Sandy Creek High School	6'9"	205	★★★★★	97	AUBURN SIGNED
6	Caleb Houstan	SF	Montverde, FL Montverde Academy	6'8"	200	★★★★★	96	MICHIGAN SIGNED
7	Michael Foster	PF	Milwaukee, WI Hillcrest Prep	6'8"	215	★★★★★	96	List
8	Kennedy Chandler	PG	Memphis, TN Sunrise Christian Academy	6'1"	170	★★★★★	95	TENNESSEE SIGNED

NBA Basketball Reference
60 Selection 2020 and Career's Stats

					Round 1			Totals				Shooting			Per Game				Advanced			
Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	3P%	FT%	MP	PTS	TRB	AST	WS	WS/48	BPM	VORP
1	1	MIN	Anthony Edwards	Georgia	1	72	2314	1392	336	211	.417	.329	.776	32.1	19.3	4.7	2.9	0.8	.017	-2.6	-0.3
2	2	GSW	James Wiseman	Memphis	1	39	836	448	226	26	.519	.316	.628	21.4	11.5	5.8	0.7	0.7	.042	-5.4	-0.7
3	3	CHO	LaMelo Ball		1	51	1469	803	302	313	.436	.352	.758	28.8	15.7	5.9	6.1	2.8	.093	1.9	1.4
4	4	CHI	Patrick Williams	Florida State	1	71	1983	655	327	99	.483	.391	.728	27.9	9.2	4.6	1.4	2.5	.060	-2.7	-0.3
5	5	CLE	Isaac Okoro	Auburn	1	67	2173	646	206	128	.420	.290	.726	32.4	9.6	3.1	1.9	0.9	.020	-5.0	-1.6
6	6	ATL	Onyeka Okongwu	USC	1	50	601	228	163	18	.644	.000	.632	12.0	4.6	3.3	0.4	1.9	.155	-1.3	0.1
7	7	DET	Killian Hayes		1	26	670	176	69	138	.353	.278	.824	25.8	6.8	2.7	5.3	-1.1	-.081	-7.3	-0.9
8	8	NYK	Obi Toppin	Dayton	1	62	683	253	138	29	.498	.306	.731	11.0	4.1	2.2	0.5	1.5	.109	-1.0	0.2

2. Data Sources

As earlier stated, we use two main datasets: ESPN recruiting database and NBA draft data.

ESPN Recruiting Database (Top-100 Rankings)

Since 2007, ESPN has ranked what it believes is the top 100 players each year. This data is widely considered as a source of truth for gauging the performance and potential of high school basketball players.

In this analysis, we consider ESPN top 100 high school basketball players from 2007 to 2015. This allows us enough years to follow and investigate players who attend some college before joining the NBA.

ESPN provides the following attributes for each top 100 player for years under consideration: ranking between 1 and 100, player's name, position, player's hometown, player's height in feet and weight in pounds, player's number of stars, player's basketball grade and the college player signed with. This is a total of 896 rows and 9 attributes.

NBA Basketball Reference Data

Basketball Reference is the go-to database for all NBA data, including draft data starting in 1947. We will be focusing on NBA drafts between 2008 and 2015 to start following 2007 ESPN top 100 cohort right after high school, and 2015 ranked players a year after college (for those players reaching NBA through March Madness, i.e. university championships).

NBA player's attributes consists of rank, team, player, college, and a whole bunch of statistics recorded through player's NBA career. With two rounds of 30 players drafted yearly, we end up with 780 observations of 22 features.

Other Data

We will find in data manipulation section that a reference of US states was required in the data clean up process.

Data Access Method

Both datasets are in HTML format accessible in URLs indicated in the table below. We used web scraping to access the tables.

In case URL formats change in the future and break, we also saved reference datasets in the csv files. The csv are provided in the zip project materials.

Title & URL(s)	Description	Variables	Years	Shape	Format	Access Method
ESPN Recruiting Database http://www.espn.com/college-sports/basketball/recruiting/playerrankings/_/order/true	High School Top 100 players per year, since 2007	Rank, Player, Position, Hometown, Height, Weight, Stars, Grade, School	2007 to 2015 (URL to be adapted in order to match given years)	896 x 9	HTML Table	Web Scraping
NBA Draft Data https://www.basketball-reference.com/draft/	NBA draft data per year, since 1947	Rank, Team, Player, College, set of career's stats (Pk, Yrs, G, MP, PTS, TRB, AST, FG%, 3P%, FT%, MP.1, PTS.1, TRB.1, AST.1, WS, WS/48, BPM, VORP)	2008 to 2020 (URL to be adapted in order to match given years)	780 x 22	HTML Table	Web Scraping
US States https://www.ssa.gov/international/coc-docs/states.html	US Two-letter state abbreviations	state, abbreviation	N/A	56 x 2	HTML Table	Web Scraping

3. Data Manipulation

ESPN Recruiting Database (ESPN Rankings)

As these rankings have high visibility as they are regularly leveraged by NBA scouts, colleges, and sports fans alike, ESPN ensures the tables are well maintained. The result is relatively clean tables, with little missing data points, misplaced fields or unnecessary addups due to the conversion from HTML.

However, the challenge was scraping multiple years of data and ensuring the scraped data was in the correct format to be put into a pandas dataframe. In order to do this, we imported the python library **requests,** which gave us the ability to use a for loop to loop through multiple years of data that lived on different web pages. As the first year of rankings was 2007, we scraped all of the years between 2007 to 2016, as players ranked after 2016 could potentially still be playing in college if they played through their senior years.

We next leveraged **regex** to scrape out extra words in columns and split out columns such as Hometown into multiple columns (Hometown, State, and HighSchool). We then ensured data quality by stripping out white spaces, updating data types that were defaulted wrong, and audited for duplicates. When looking for duplicates, we found 6 players that appeared in the dataset twice. Even though some of those players were found to be homonyms instead of duplicates by looking at other data points other than name, we dropped them from the data set to avoid confusions in merging with NBA draft dataset that does not have the same attributes.

Checking on data accuracy, we also noticed that unique states include Quebec and Ontario Canadian provinces which have NBA teams. We also found some overseas countries or continents, more difficult to filter. The additional abbreviations states dataset from US gov help to build consistency around US states.

Regarding missing data, we decided not to drop players missing some attributes, since we anticipated that ESPN and NBA tables as far as we had a player name.

Basketball Reference Data

As the Basketball Reference dataset was in a similar table structure as the ESPN data set, we were able to again leverage the request library to loop through the data to scrape years 2008-2021 of data. As there are two rounds in the NBA draft, one webpage included 2 tables representing both rounds for that year. This created a challenge when looping through the years as it would duplicate table column names when scraping for the 2nd table. In order to avoid creating these column names as a row of data in our data frame, we found the lines numbers that represented these extra column names and dropped them when looping through.

Once we converted the years into a dataframe, we followed similar steps to the ESPN data set to clean and validate data accuracy.

Merging Datasets

Going back to the main reason for our analysis of trying to determine if the ESPN ranking data is a good predictor of a high school basketball player making it to the NBA or not, we decided to left join the Basketball Reference dataset onto the ESPN dataset by names. This way, no players in the ESPN dataset would be dropped, and it would line up the draft data if they were drafted. This new master dataset contained all the high school players ESPN ranking data and signaled whether or not they actually got drafted to the NBA, in what year if the did, and with what pick. Once merged, we checked the shape of the new master dataframe to ensure no rows were dropped from the ESPN data set. The number of rows were identical.

Cleaning the Master Data Set

Now that we had all of the data merged together, we needed to go through and clean up the data set by striping out blank spaces and ensuring columns were the correct datatype. As the height column was pulled in as a sting (e.g. 6’6”) representing feet and inches, we needed to convert this into an integer. We did this by splitting out the column into a feet and inches column and by leveraging regex to drop any left over punctuation. Next by dropping additional white spaces in these columns and converting the columns into integers, we were able to perform calculations to get a height in inches column.

Lastly, as the player’ position column was represented by letters (e.g. ‘C for center), we created a dictionary that linked the letter to a number to represent the position as an integer and created a new column in the data set to reflect that number by leveraging the map function. With these new integer columns built from strings, we could now leverage these columns to feed into our model.

4. Preliminary Analysis

Subset considered for the analysis

Most NBA draft attributes are beyond the scope of this study, more specifically all columns on career' statistics. We further sliced the master dataset accordingly, to only keep ESPN like features in addition to ESPN ranking year, NBA drafting year and a Boolean variable indicating whether a player under consideration was drafted. The left table below gives a sample observations of the analysis subset with the selected attributes.

Sizing up the Master Dataset

As we begin looking through the data set to get more familiar with the columns and the layout of the data, we leveraged **describe()** to give us some insights into the high-level statistics of the data. The table at bottom-right summarizes those simple statistics.

One of the takeaways from this view that was helpful for us is the min grade a player can receive by ESPN is 80. In other rankings, an 80 could be perceived as a relatively high score, but in this situation, it signals a player at a lower rank.

Other interesting takeaways that stood out is with the height column. The tallest player listed in the ESPN rankings measured in at 89 inches which is equivalent to 7 feet 5 inches. The shortest player stands at only 66 inches which is equivalent to 5 feet 6 inches. The median height is 78 inches or 6 feet 6 inches signaling that although the disparity is high between the min and max, height is likely going to be a major factor in the ESPN ranking system.

As we look at the other columns in the table, we can believe that as well will influence a player getting ranked in the top 100 and believe that these variables could also influence whether or not a player makes it to the NBA. These view gives a good idea of the data and allows us to better understand the data as we continue to build out additional views.

Sample Observations for the final master dataset

	ESPN Rank	Player	Position	Position_number	Height_inches	Weight_pounds	Grade	ESPN_Ranking_Year	Hometown	State	NBA_Drafting_Year	Drafted
556	68	Xavier Johnson	SF	3	78	185	93	2012	Temecula	CA	NaN	0
213	20	Alex Oriakhi	C	5	80	230	96	2009	Lowell	MA	2013.0	1
483	96	Kevin Ware	SG	2	77	180	92	2011	Conyers	GA	NaN	0
313	21	Joe Jackson	PG	1	72	175	96	2010	Memphis	TN	NaN	0
437	49	Ben McLemore	SF	3	77	185	95	2011	Saint Louis	MO	2013.0	1
544	56	Georges Niang	PF	4	79	230	94	2012	Methuen	MA	2016.0	1

Simple Statistics

	ESPN Rank	Position_number	Height_inches	Weight_pounds	Grade	ESPN_Ranking_Year	NBA_Drafting_Year	Drafted
count	884.00	884.00	884.00	884.00	884.00	884.00	229.00	884.00
mean	49.99	2.91	77.92	201.37	92.14	2011.02	2013.29	0.26
std	28.61	1.31	3.25	25.92	4.62	2.58	2.71	0.44
min	1.00	1.00	66.00	150.00	80.00	2007.00	2008.00	0.00
25%	25.00	2.00	76.00	180.00	89.00	2009.00	2011.00	0.00
50%	50.00	3.00	78.00	200.00	93.00	2011.00	2013.00	0.00
75%	75.00	4.00	80.00	215.00	96.00	2013.00	2015.00	1.00
max	100.00	5.00	89.00	330.00	99.00	2015.00	2019.00	1.00

5. Descriptive Analysis [Choropleth Visualization]

Choropleth of US States Representation In ESPN Rankings

As location was not a variable that was included in the describe statistics, we decided to build a visual to see if there is any relationship between where a player is from and whether or not they are ranked. In order to do this, we leveraged a Choropleth visual from the **plotly** library to visually highlight states that produce the most ESPN ranked players as well as players that end up getting drafted. In order to do this, we first had to group the data by states to pass it into the map. The map is dynamic so when you hover over a certain state, it creates a pop-out for you to see both the number od ranked players from that state as well as the number of those ranked players that were drafted.

From this view, we see there is a large number of ranked players from California, Texas, Georgia, and Illinois. As these states are also states that have high populations, in future projects, we would like to pull in state population to see which states produce the most ranked and drafted players when controlling for state population.

Choropleth for home state of ESPN Top-100

Time Series Visualizations for further Insights

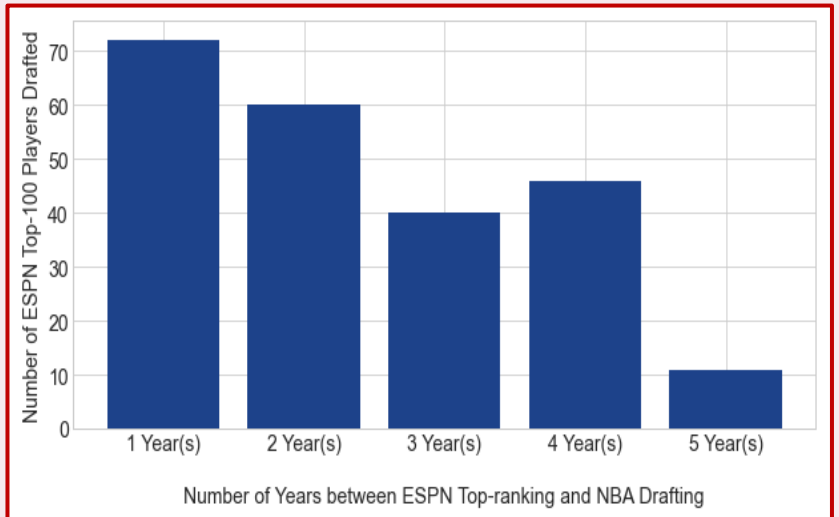
The next slide presents a set of histogram and time series visualizations, generated using **matplotlib** library.

The histogram shows the number of years it takes on average for a player ranked by ESPN to be drafted by the NBA. We see that the great majority of players join the NBA in their first or second year after high school. Numbers drafted after 3 vs 4 years indicate that players not drafted within the first couple of years prefer completing their Bachelor's degree before moving onto their NBA career.

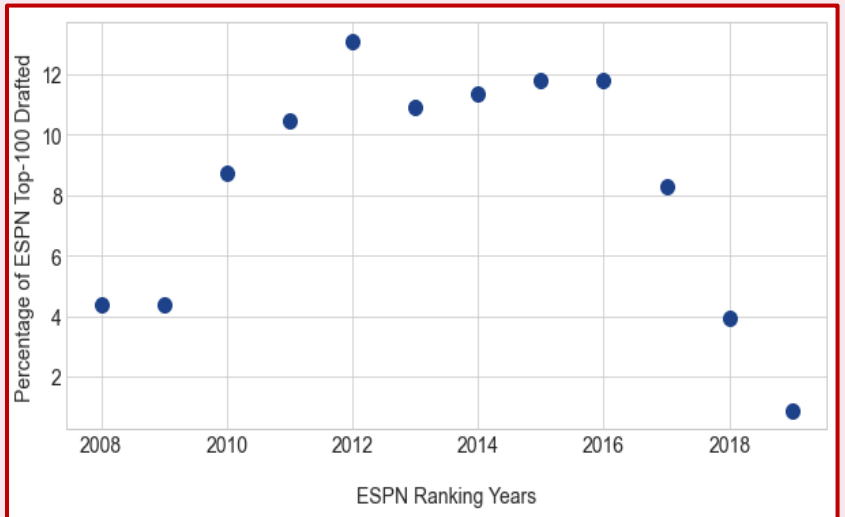
The following time series provides percentages of ESPN top ranked players drafted into the NBA over the years. Percentages range from 8 to 12% from 2010 to 2017, which might be considered as average. Low percentages before 2010 and after 2017 are explained by the fact that players ranked before 2007 and after 2015 are not considered in the study.

The other time series plots compare different attributes for drafted vs non-drafted players. More specifically, we observe higher average grade for drafted players. Average ranking is also much better(i.e., low) for drafted players. Regarding height and weight characteristics, drafted players are taller than undrafted players. Except in 2008, we also noticed a higher average weight for drafted players.

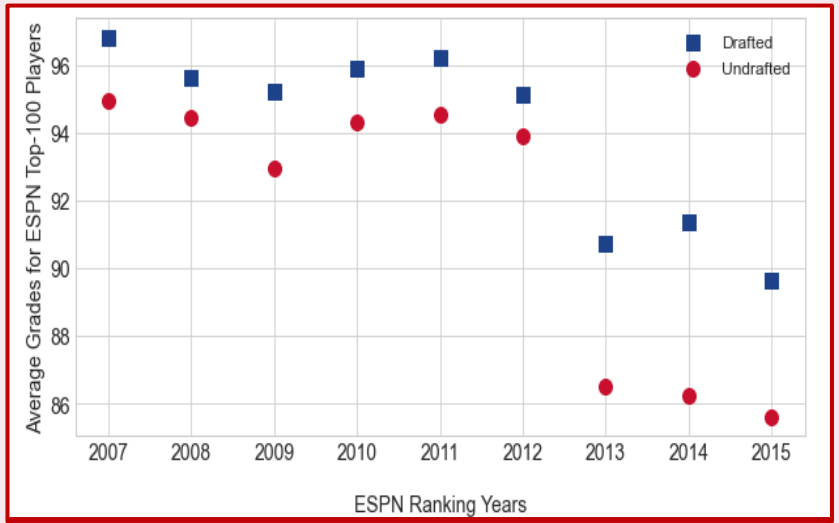
Descriptive Analysis [Time Series Visualizations]



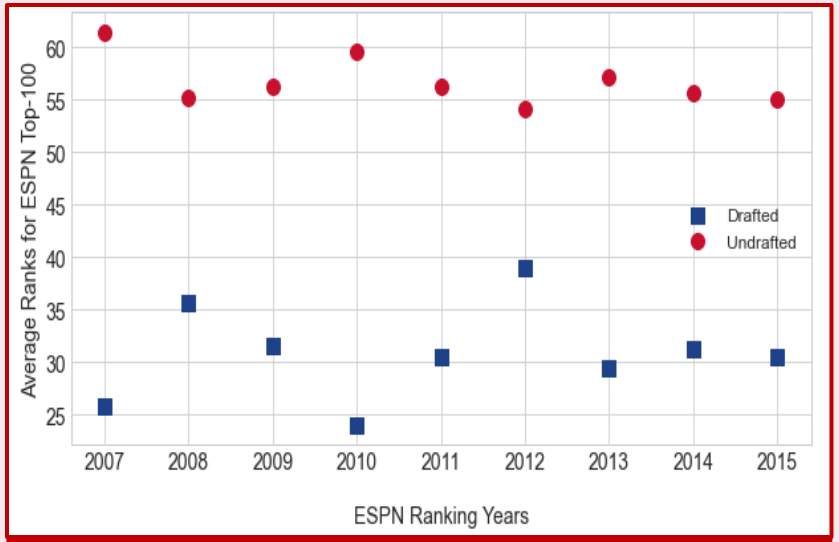
Number of Years from ESPN Ranking to NBA Drafting



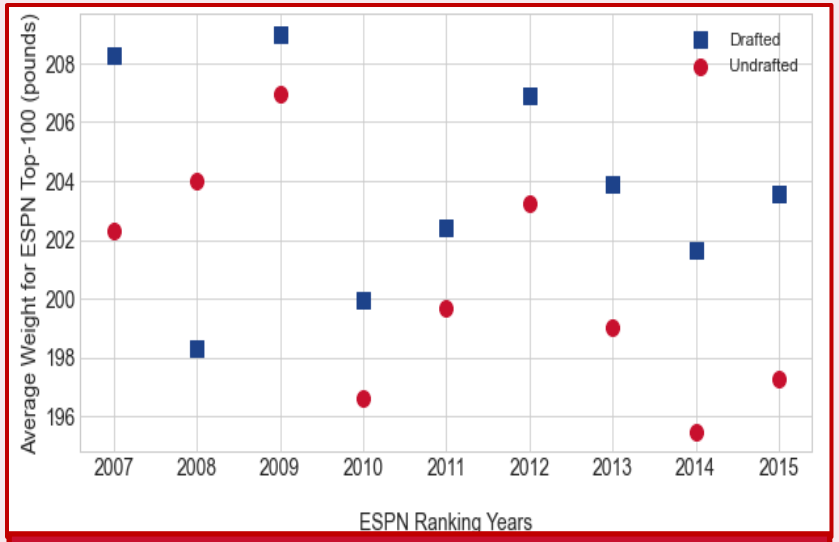
Time Series for % of ESPN Top Players Drafted in NBA



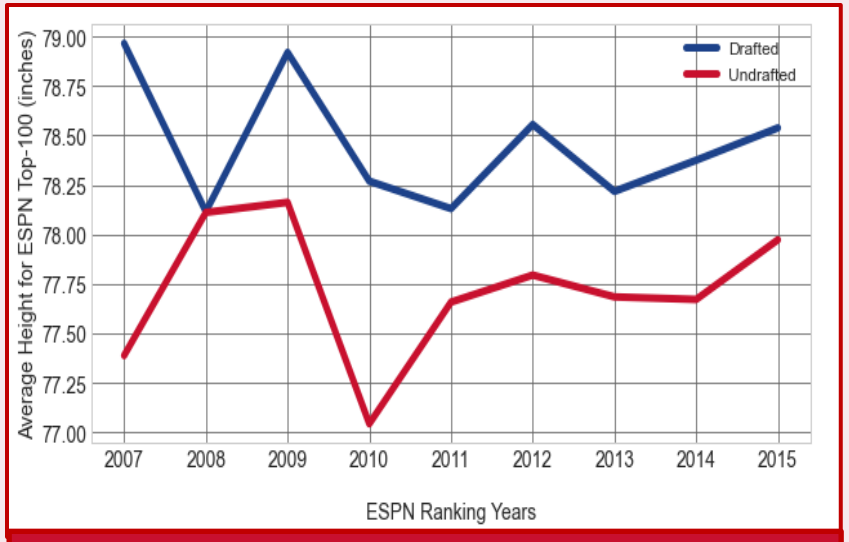
Time Series for Average Grades of ESPN Top Players



Time Series for Average ESPN Top-100 Player's Ranks



Time Series for Average Weight of ESPN Top Players



Time Series for Average Height of ESPN Top Players

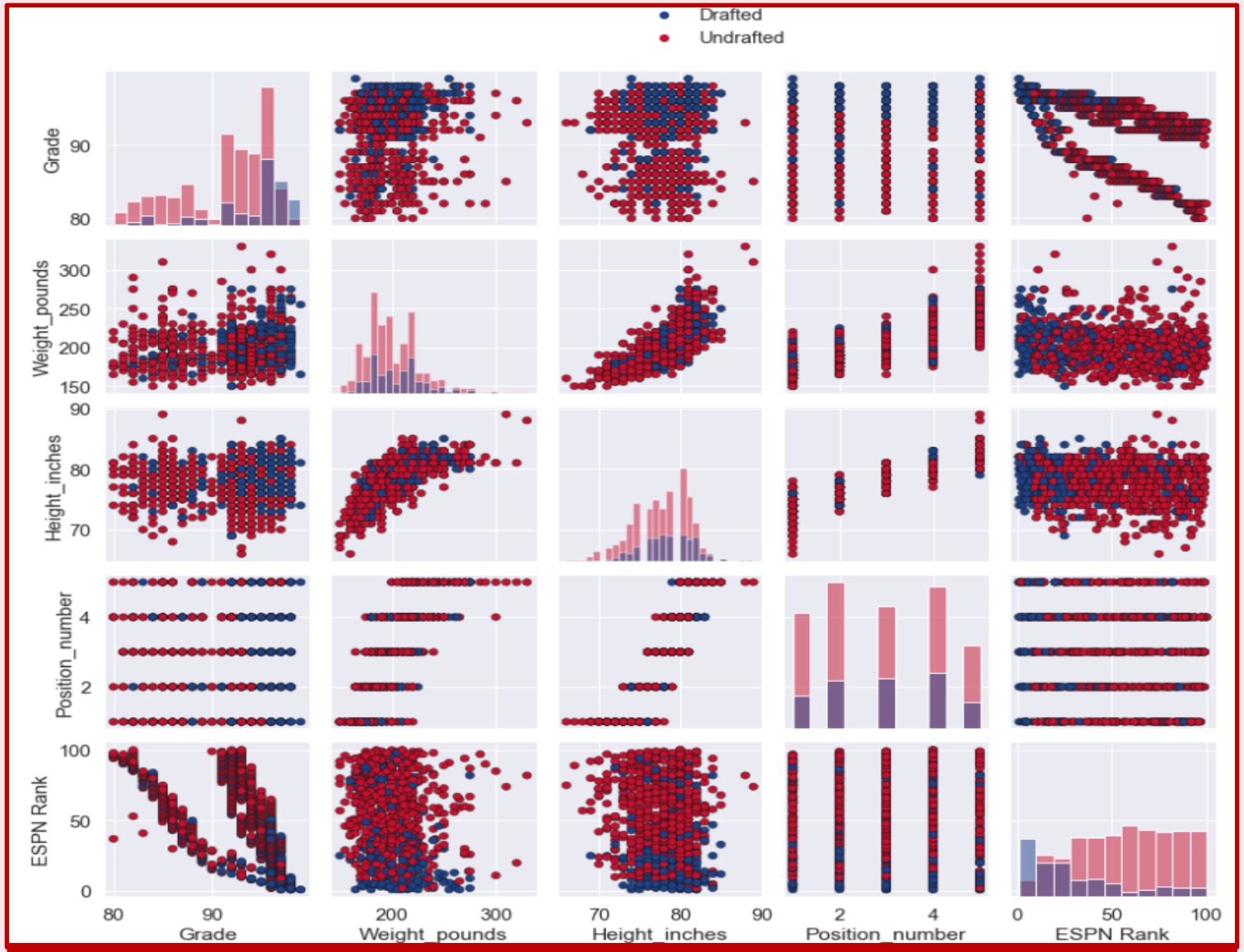
Descriptive Analysis [SPLOM Visualization]

SPLOM

Building off of our describe statistics, we decided to build a SPLOM view to see if we could begin to visually see any correlations between our data. In order to build a SPLOM visual, we leveraged the **seaborn** library.

One advantage of the SPLOM visual for this analysis is that we can represent both drafted and undrafted players by different colors. This helps us to see any potential differences in correlations between players that were ranked and drafted vs non-drafted.

When examining the different charts, its becomes obvious that characteristics such as grade, weight, height, and player rank all are positively correlated with getting drafted. For example, in the bottom left “Grade” column, we can see groups of blue near the right of all the charts in the columns signaling that higher ESPN grades do have a factor in whether or not you end up getting drafted into the NBA. It does also show outliers as there is blue (drafted) player dates near the left of the charts in the “Grade” column as well. This SPLOM visual helps us see that these variables are important variable that influence whether or not a players gets drafted or not, but we still are unable to tell how accurate. This visual helps lay the foundation for us to continue to analysis the data leveraging an ML model.



SPLOM for Correlation between Key Player’s Features

6. Predictive Analysis [KNN Supervised Learning]

How predictive is ESPN rank on performance?

As gauging the predictability of the ESPN ranking system is a binary problem (either yes or no), we decided to leverage a K-Nearest Neighbor (KNN) model from the sklearn library as this is proven model for classification problems.

When first testing the model with just the Grade variable, it appears that the model produces around a 71% accuracy for predicting whether or not a player will make it to the NBA based off of their ESPN Grade. When we add additional variables that appeared to be heavily correlated with players who get drafted based off of the SPLOM visual, we see that the accuracy of the model jumps to 79%.

In order to best gauge which n_neighbors produces the most accurate prediction with these multiple variables, we created a for loop to gauge which n_neighbors tunes the model for further accuracy (figure 2). Based off of this tuning, 13 produced the best accuracy.

Finally, in order to evaluate the model to see what the prediction would be based off of the mean scores produced by the describe analysis previously discussed, we plugged those inputs into the model to get the answer of 0 (figure 3). This signals that if you are an average player in the ESPN rankings with average qualities such as height, weights, etc… compared to your peers, you will likely not make it to the NBA.

```
X = master_df['Grade']
X = X.values.reshape((-1,1))
y = master_df['Drafted']

knn = KNeighborsClassifier(n_neighbors = 5)

# default is 75% / 25% train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
knn.fit(X_train, y_train)
knn.score(X_test, y_test)
```

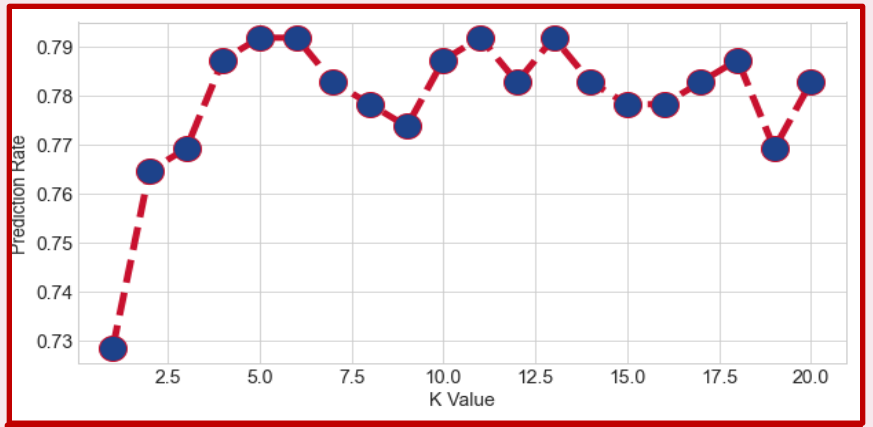
```
0.7104072398190046
```

```
X = master_df[['Grade','Weight_pounds','Height_inches','Position_number', 'ESPN Rank' ]]
y = master_df['Drafted']

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
#y_train.shape
knn = KNeighborsClassifier(n_neighbors = 13)
knn.fit(X_train, y_train)
knn.score(X_test, y_test)
```

```
0.7918552036199095
```

1) Train



2) Optimize

```
prediction = knn.predict([[92,201,78,3,50]])
prediction
```

```
array([0])
```

3) Evaluate

7. Further Discussion & Statement of Work

Conclusion

It appears that from the dataset we were able to create from the ESPN ranking and NBA draft data, that there are distinct patterns of characteristics that players have that are correlated with making it the NBA. Out of the over 500k players that ESPN could choose to include on their yearly high school basketball rankings, it does appear that they are able to pinpoint players that have a high probability of making it to the NBA. The higher the grade and rank you receive, the higher the probability you will make it.

Possible Future Investigations

As we think about potential next steps for this analysis, we feel it would be beneficial to pull in additional data into the analysis to get additional perspective. Data that we would want to pull include ranking of the college basketball program a player chooses to attend, college stats, and data around any potential injuries that are sustained during college. These additional data points would give increased visibility into whether or not a player can help or hurt their cause in college compared to their previous high school player rankings. This additional information could be valuable for players trying to determine which schools are realistic for them to attend based off of their current skill level. This would also help colleges be able to focus on athletes that best fit their program so they can spend the most time recruiting those players. Also, if you're are a late bloomer whether from a physical or skill perspective, it would also help you be able to gauge what goals you should set to propel yourself to the next level.

Statement of Work

Diane researched the topic, wrote the code to be able to scrape the data from the web, built data frames from the web scrapes, cleaned and manipulated the data, merged the data sets, built the Choropleth visual, the SPLOM visual, several scatter plots, merged both jupyter notebooks together to be able to turn in 1 notebook, and reviewed/edited the report.

Tanner researched the topic, wrote the project proposal ,identified the data sources and how they could be combined, helped clean and manipulate the data, built histograms, the describe analysis, the KNN model and analysis, and did the first draft of building the report.

We both met multiple times a week for the last 6 weeks to collaborate, determine priorities, and help each other throughout the process.

Overall, we enjoyed doing this basketball analysis and had great team collaboration. The only visualization that we were interested to include but couldn't get work is the representation of decision boundaries for our KNN algorithm. Decision boundaries for 3-D or plus representation might be challenging, as is the case (i.e., five) with our training set.