Finding Breakout Wide Receivers in Fantasy Football

1.) **Description**

Fantasy football has become one of the most popular products of the National Football League (NFL) over the past twenty years. In fantasy football, participants become managers and draft imaginary team rosters by selecting players from current, real-life NFL rosters. Every week, managers face off with one another with their teams by setting a lineup of players who score points based on their actual performance in real-world games. The team scoring the most points at the end of the week is awarded a win and the other team a loss. The most common drafting system is a snake draft, where the draft order is reversed every round until all rosters are full. Each pick has a value, with earlier picks being more valuable than later picks. This is because as the pool of players gets smaller, managers become less confident about which players will lead to fantasy success. By compiling the results of drafts, each player's average draft position (ADP) is calculated, which can be considered a proxy for who the public thinks are the most dominant fantasy players in descending order. Typical fantasy leagues feature 12 teams with 16 rounds in the draft. Finding players with higher upside (who will score more than expected) in the draft's middle rounds is one of many possible ways to build a successful fantasy roster. This project adopts this approach and attempts to identify which wide receivers (position of player) for the 2023-2024 NFL season have the potential of outscoring their ADP. Such players will be called breakout wide receivers. For this project, breakout wide receiver will be defined as any wide receiver with an ADP in rounds 6-9 (12 team league; pick #: 60-108) who outscores their projected fantasy points per game by 'X' number of points. A simple linear regression model will be created to identify breakout wide receivers using data from the NFL's past ten years (2013-2022). Insights gained from that model will be then applied to the 2023 data to identify potential breakout wide receivers for the 2023-2024 fantasy football season.

2.) **Data**

a. and **b.**

The data was collected from three sources utilizing the libraries of requests, BeautifulSoup, re, and pandas. The first source scraped was Pro-Football-Reference's yearly fantasy rankings, which contain a ranked table of each NFL player who scored any number of fantasy points based on their real-life season player stats. The second source scraped was myfantasyleague.com, a site that contains historical ADP data with filters that can be adjusted to league settings. The last source scraped was drafthistory.com for its data on the yearly, actual NFL draft for later feature engineering.

```
source_1_URL = f'https://www.pro-football-reference.com/years/{year}/fantasy.htm'
source_2_URL = f'https://www46.myfantasyleague.com/{year}/reports?R=ADP&POS=*&PERIOD=AUG15&CUTOFF=5&FCOUNT=12&ROOKIES=0&INJURED=0&IS_PPR=2&IS_
source_3_URL = "https://www.drafthistory.com/positions/wr.html"
```

The first source contained a table with two header rows containing relevant column data. The page source was inspected to determine the element type and id of the table ('table' and 'fantasy' respectively). Code was written using requests and BeautifulSoup by creating a soup object of the parsed HTML content and searching for the table of interest by specifying the target element type and id. Simple cleaning of the table column names,

such as accounting for the information provided by the two header rows, was performed. Symbols were removed from all player names for easier merging of data sources. Only relevant information from the table was saved to a pandas dataframe. These steps were performed in a loop to scrape through multiple years of data. The data was saved within each loop as a csv with appropriate naming conventions. Lastly, all the scraped data was saved to a csv as a master copy of the data for easier access during cleaning and analysis.

The second data source mainly employed the same scraping technique. First, the settings were adjusted on the actual web page to generate the appropriate URL for the league settings of this project. Upon inspection of the HTML content, the target element type and class were determined to be 'table' and 'report.' This site was also scraped utilizing a loop to collect ADP data from the past ten seasons. One issue I ran into was reformatting the information in the 'PLAYER' column. It contained the player's name (last name, first name), the team they play on, and positional information. The information in this column had to be parsed and edited to match the formatting of previously saved data. A function called name_splitter was created with the regular expressions library (re) to take the input of the 'PLAYER' column and return a string formatted as "first_name last_name." With the appropriate initial cleaning applied, ADP by data by season and a master copy was saved.

The third data source did not require a loop ,as only one page had to be scraped. After identifying the table utilizing request and BeautifulSoup, the first row of the table was eliminated and the second row was set to the column names. The presence of empty table cells was fixed by forward-filling the year column. This table was then saved like the previous sources.

The total number of data samples collected would amount to the sum of the three data sources. The first data source generated a dataframe with 6151 rows by 22 columns. The second and third data sources generated dataframes with 3,831 rows by 6 columns and 2,286 rows by 8 columns, respectively. Considering the data samples scraped for the column names (thus adding "1" to each row count), 176,626 data points were scraped.

3.) **Analysis and Visualizations**

a. and **b.**

After cleaning, pre-processing, and feature engineering on the raw data files (saved as csv in the /data/processed folder), all three data sources were merged using inner joins on the appropriate columns filtered for only wide receivers creating the analysis dataframe (saved as wr_analysis.csv) with the following as columns:

Column Name	Column Name	Column Name
"Player"	"Season"	"Tm"
"Pos"	"Age"	"G"
"GS"	"Tgt"	"Rec"
"PassingYds"	"PassingTD"	"PassingAtt"
"RushingYds"	"RushingTD"	"RushingAtt"
"ReceivingYds"	"ReceivingTD"	"Int"
"Fumbles"	"FumblesLost"	"PPR"
"Rank"	"Avg Pick"	"rookie_season"
"No."	"Round"	"Pick"
"College"	"current_yr"	"FPPG"

The "G" column was subjected to panda's df['col'].value_counts(normalize = True) to get the relative frequencies for each unique value. This provides a breakdown of how many games each wide receiver played in a given season. Those with smaller sample sizes (i.e., fewer games played in a season) can be possible outliers. Thus, all those with less than nine games played (approximately 10.7% of the analysis dataset) were removed from the analysis. To establish the viability of creating a logistic regression model, ADP was plotted against FPPG using seaborn's lmplot; this function also fits a regression model across the plotted data with the model's equation and R^2 (Figure 1). The R-squared value of 0.43 indicates that the model does a decent job of modeling FPPG as a function of ADP; the coefficient of the x variable shows that there is a negative correlation between the target and instance variables. Intuitively, this makes sense as players expected to score more points over the season will likely be selected before those less likely to score points.

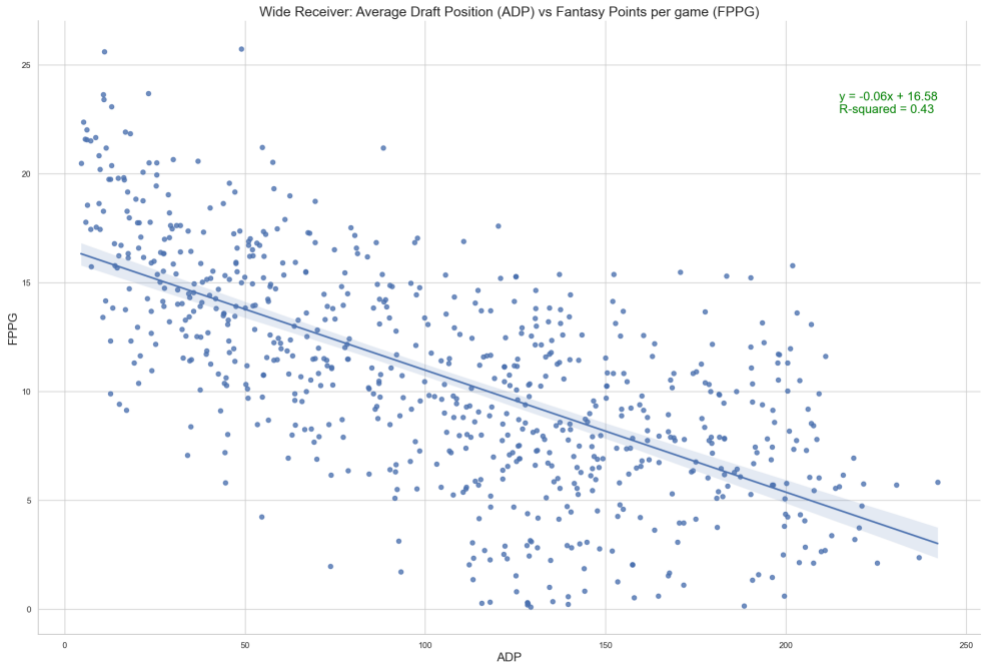

Figure 1. Scatter plot of ADP vs FPPG fit with a regression model using seaborn's lmplot()

Using the above feature and target variables as starting points, a regression model was created using the ordinary least squares method utilizing the statsmodels API. Multiple models were investigated. The one with the highest R^2 score and statistically significant feature variables was FPPG as a function of ADP and ADP^2 (Table 1). This iteration of the regression model was used to predict every player's FPPG for a given season as a function of ADP using statsmodels. The predicted and actual FPPG were then plotted, demonstrating a better fit than our original model (Figure 2).

OLS Regression Results

Dep. Variable:	FPPG	R-squared:	0.474
Model:	OLS	Adj. R-squared:	0.473
Method:	Least Squares	F-statistic:	328.2
Date:	Tue, 05 Dec 2023	Prob (F-statistic):	2.77e-102
Time:	13:10:42	Log-Likelihood:	-1977.5
No. Observations:	730	AIC:	3961.
Df Residuals:	727	BIC:	3975.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	19.1246	0.428	44.716	0.000	18.285	19.964
ADP	-0.1234	0.009	-13.763	0.000	-0.141	-0.106
np.power(ADP, 2)	0.0003	4.03e-05	7.779	0.000	0.000	0.000

Omnibus:	2.149	Durbin-Watson:	1.639
Prob(Omnibus):	0.341	Jarque-Bera (JB):	2.048
Skew:	-0.067	Prob(JB):	0.359
Kurtosis:	2.778	Cond. No.	6.18e+04

Table 1. Ordinary least squares regression model results for FPPG as a function of ADP

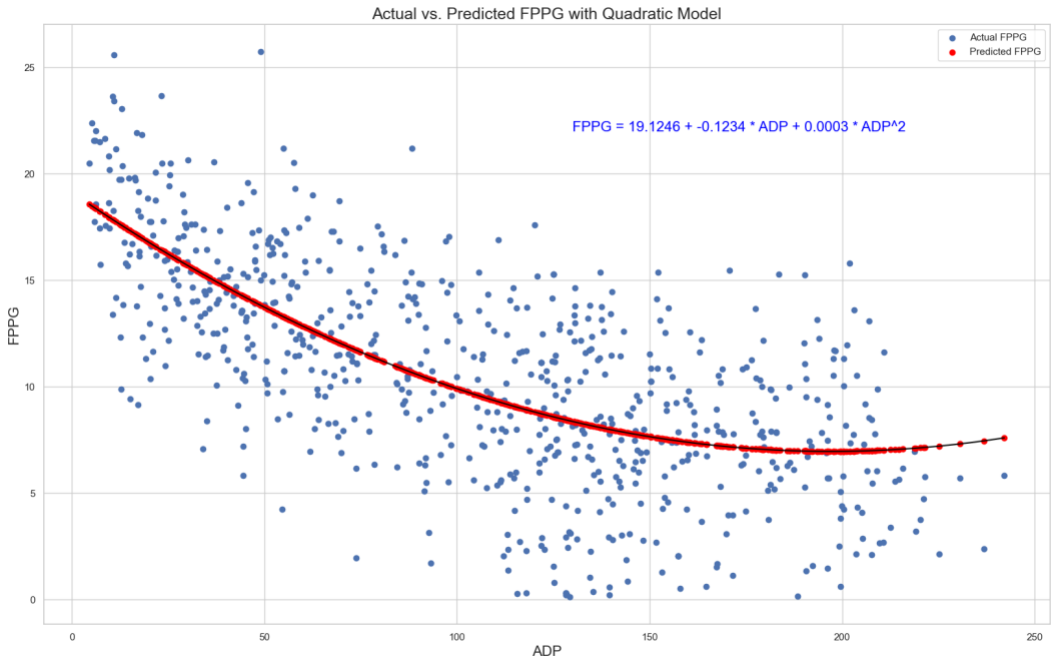

Figure 2. Scatter plot of the actual and predicted FPPG that were modeled as a quadratic equation of ADP

Because the focus of this project is on rounds 6-9 of a 12-team snake draft, those with ADP between 60 and 108 inclusive were filtered for and a new column ("diff_ppg") for the difference in FPPG between the actual and predicted was added to the latest analysis dataframe (named r_mid_wr). A multi-stack histogram of this new feature was created with the current number of years in the NFL as the hue (Figure 3).

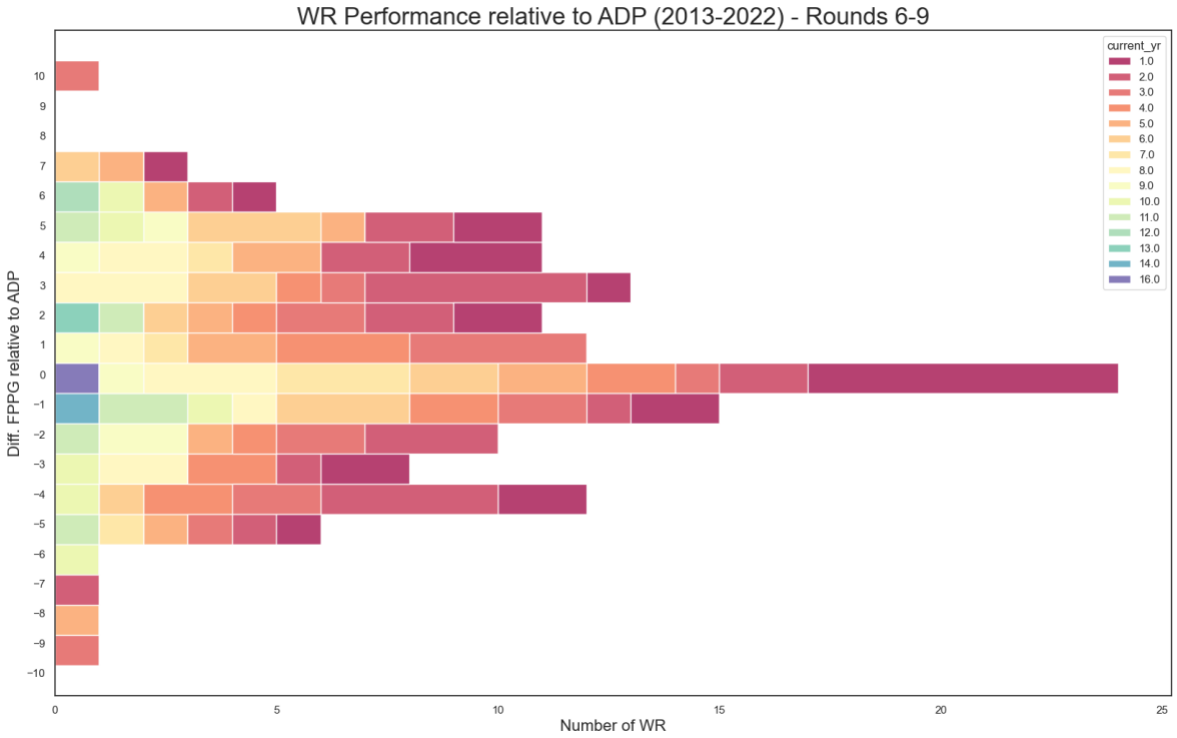

Figure 3. Wide receiver performance defined as the difference in FPPG relative to ADP histogram, with the current number of years in the NFL set as the hue

As this project aims to find breakout wide receivers, the players of interest have a positive difference in FPPG relative to the predicted value. We observe a slight spike at the bin of +3. This value was used to arbitrarily set our "X" value to define a breakout wide receiver as a player who outscores their predicted FPPG by greater than or equal to "3"; the analysis dataframe was further filtered to only include those based on such parameters. Based on this filtered dataframe, two bar charts (distribution charts) were generated based on current year in league and NFL draft round capital (Figures 4 and 5).

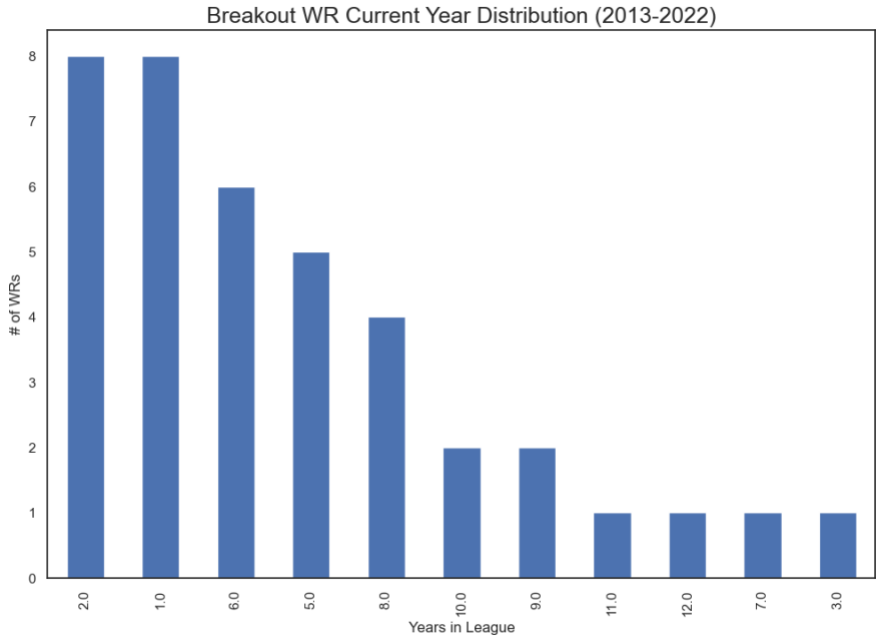

Figure 4. Breakout wide receiver distribution by number of years in the NFL

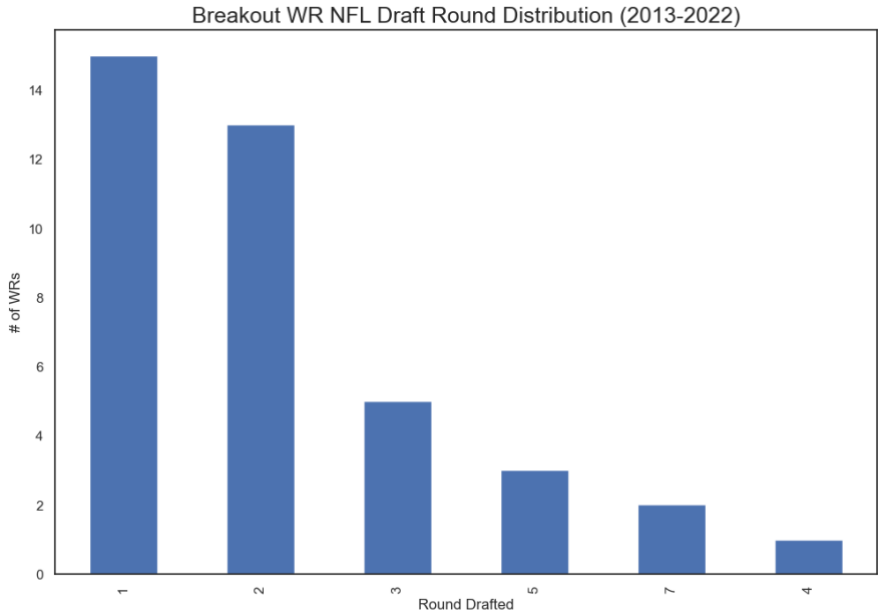

Figure 5. Breakout wide receiver distribution by NFL draft round capital

c. and **d.**

Figure 4 shows that 56% (22 out of 39) of breakout wide receiver seasons occurred in years 1, 2, or 6. Figure 5 shows that 72% (28 out of 39) were drafted in rounds 1 or 2 of the NFL draft. This makes some logical sense, as NFL teams are investing more draft capital for a wide receiver with earlier-round picks. We expect such players to get on the field in their rookie season and likely improve in their second year. Such findings were applied to the 2023 ADP data of wide receivers drafted in the middle rounds to predict possible breakout candidates (Table 2).

Rank	Player	ADP	Pos	Tm	Season	rookie_season	No.	Round	Pick	College	current_yr
61	Christian Kirk	62.19	WR	JAC	2023	2018.0	5	2	15	Texas A&M	6.0
72	George Pickens	71.21	WR	PIT	2023	2022.0	11	2	20	Georgia	2.0
76	Jordan Addison	73.18	WR	MIN	2023	2023.0	4	1	23	USC	1.0
81	Jaxon Smith-Njigba	78.69	WR	SEA	2023	2023.0	1	1	20	Ohio State	1.0
82	Jahan Dotson	78.96	WR	WAS	2023	2022.0	5	1	16	Penn State	2.0
83	Zay Flowers	79.60	WR	BAL	2023	2023.0	3	1	22	Boston College	1.0
89	Courtland Sutton	83.90	WR	DEN	2023	2018.0	3	2	8	Southern Methodist	6.0
108	Quentin Johnston	98.65	WR	LAC	2023	2023.0	2	1	21	Texas Christian	1.0
128	Skyy Moore	105.96	WR	KCC	2023	2022.0	13	2	22	Western Michigan	2.0

Table 2. Analysis output of potential breakout wide receivers for the 2023-2024 fantasy football season

In conclusion, we created a model for FPPG and applied that to our scraped data to find potential breakout wide receivers for the 2023-2024 football season. The regression analysis indicated a negative correlation between ADP and FPPG, with an R^2 value of 0.43. This suggests that as the draft pick number increases (i.e., indicating a later pick), the FPPG tends to decrease, which aligns with intuitive expectations. Characteristics of such wide receivers were found by applying the most effective model in predicting FPPG. These findings give us insight into players to target in drafts.

The impact of such findings is the further extension of knowledge to the fantasy football world. Understanding ADP, FPPG, and actual player performance helps fantasy football managers make better-informed decisions during drafts. The insight into potential breakout wide receivers' characteristics can help managers find undervalued players, one of many key components for a successful fantasy season. The revealed insights are a foundation for further exploration and analysis into drafting strategies and more predictive and accurate model building.

4.) **Future Work**

Given more time, supervised feature selection would have been employed to enhance the identification of the most relevant and significant variables for analysis. Doing such is a much more efficient process of model building, thus facilitating improved model refinement. Investigating multivariate and more complex polynomial equations could help capture complex, unseen relationships, elevating the model's precision. Furthermore, subjecting the model to further training and testing will help us better assess its generalizability. These future directions will provide even deeper insight into fantasy football, advancing the pursuit of a more comprehensive understanding.