EXECUTIVE SUMMARY

PLAYER MANAGEMENT: AN ANALYSIS OF NBA PLAYERS AND THEIR CONTACT VALUES

DAMOND ALLEN & ADRIANNA HIGH | WINTER 2024

Background

Los Angeles-based player management firm wants help identifying target players for their recruiting efforts and areas to help current and future clients maximize their contract value. Additionally, when considering new clients, the firm wants to know how likely the player is to be offered a top-value contract.

Objectives

- Investigate the relationship between player contract value and their professional and college performance stats.
- Identify metrics that contribute to a large contract value
- Likelihood of top-value contract

Data and Cleaning Methods

- *Player Contract Value:* retrieved from an NBA stats website, manually downloaded as a .xls file, renamed, and saved as a .xlsx file. The critical variables are the 2023-2024 contract value and player names; the remaining columns are dropped.
- *Player IDs:* Retrieved using an API, the important variables are player name and player ID. This dataset required some manual manipulation of names to allow for joining to contract value dataset.
- *Player career stats:* Retrieved using an API, there are several essential variables, including points, steals, assists, and blocks, to name a few. This dataset contains season totals for each season, and the metrics were aggregated to averages per game.
- *NBA Teams*: The NBA API is used to extract team IDs, and the data is returned as a dictionary with only the team ID.
- *Draft History*: The NBA API extracts the draft history of all 30 NBA teams. The essential variables are draft round, pick, and college. This dataset was filtered only to include players drafted from a college. We also pre-processed the names by removing punctuation.
- *College Stats*: College statistics were extracted for each player from their player profile page on Sports Reference. These webpages were stored locally and parsed to isolate the Players college career totals

Interesting Insights

- A moderately strong positive correlation exists between turnovers and a player's contract value. However, this case of correlation does not equal causation because turnovers hurt the game. It is not recommended that players increase their turnover rates.
- The median games played across the 419 players in our dataset is just 206 games or 2.5 seasons. Players making more than $10 million per year tend to play more than 250 games over a 5-year period.

Outcome

- Increasing average field goals by 1 **adds nearly 9 million dollars to contract value** when controlling for other factors.
- Increasing average fouls by 1 **subtracts 9.8 million dollars from contract value** when controlling for other factors.
- Increasing average assists by 1 **adds 1.3 million dollars to contract value** when controlling for other factors.
- New Mexico State university players tend to have higher contract values.

MOTIVIATION

To Help Players and Agents Capitalize On A Growing NBA Market

Background

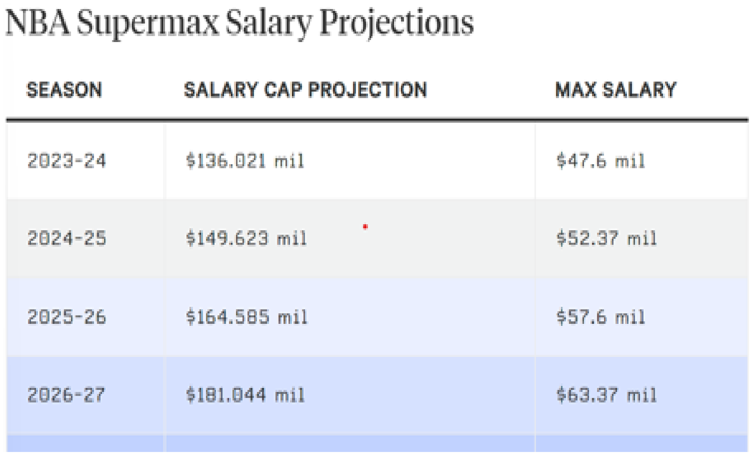
For this project, we have created a hypothetical scenario in which an agent of a Los Angles player management firm hired us to explore the leading factors impacting an NBA player's guaranteed contract value. The agent receives a 5% commission on each contract he negotiates; the higher the value, the more money he earns. The agent will use the analysis to advise players and prioritize performance improvement metrics.

Context

Like most other sports, the NBA sets a salary cap that determines how much a team can spend on salaries. Since some franchises reside in more significant, popular cities, they generate more revenue than smaller market teams. The cap aims to keep the game competitive by limiting the amount of money big market teams can spend on their roster of players. The cap is partly determined by the league's total revenue, which has doubled in the last decade. As the NBA makes more money, the players and their respective agents stand to gain financially. Figure 1. on the right shows salary projections reaching more than 80 million annually by 2029.

Similar Studies

An article titled "NBA Player Salary Analysis based on Multivariate Regression Analysis" was published by Highlights in Science, Engineering and Technology. The authors conducted a similar study but focused on ways to balance disproportionate salaries between the players. This study gave us an idea of what variables to investigate for our project. Additionally, we referred to a prior study from Koki Ando, who also conducted a regression analysis, but his goal was to predict a player's future salary; from that study, we got an idea of how to structure our experiment.



NBA Supermax Salary Projections

SEASON	SALARY CAP PROJECTION	MAX SALARY
2023-24	$136.021 mil	$47.6 mil
2024-25	$149.623 mil	$52.37 mil
2025-26	$164.585 mil	$57.6 mil
2026-27	$181.044 mil	$63.37 mil

Figure 1. Table of supermax salary projections. *NBA salaries keep going up. Prepare to have your mind blown in the future* by Mike Vorkunov, 2023, The Athletic. https://theathletic.com/4740069/2023/08/03/nba-salary-cap-rise-jaylen-brown/

Goals

For this project we aim to deliver the following to the agent:

- A specific list of factors likely to increase or decrease contract value
- Expected contract values for newly signed players based on college performance
- A list of schools to focus recruitment efforts

DATA SOURCES

Name	Description & Important Variables	Size	Format	Links
Contract Values	The excel file was downloaded from Basket Reference. It contains: Rank, Team, and Guaranteed Total Salary. For our analysis, we selected **the 2023-24 salary** to represent the player's contract value.	42 KB	Microsoft Excel	2023-24 NBA Player Contracts
Player IDs	The players ID data frame extracted from the NBA API was necessary to merge contract values with NBA career stats. It contains only three columns: **Player's First Name, Player's Last Name and Player's ID.** These data are apart of the API's static library.	99 KB	Pandas DF	API Documentation
Player Career Stats	Player career stats were extracted from the NBA API. It provides the season level statistics for all active players for the last 5 seasons. Important variables include: **Games (Played & Started), Field Goals (2-point & 3-point), Free Throws, Offensive & Defensive Rebounds, Assists, Steals, Blocks, Turnovers and Points.**	419 KB	Pandas DF	API Documentation
NBA Teams	The NBA teams dictionary contains NBA team specific variables: **Team ID, Full Name, City, State and Year Founded**. We **only extracted the team ID** to use as input for API call to request team draft history. These data are apart of the API's static library.	312 B	Pandas Dictionary	API Documentation
Draft History	Draft history for each of the 30 active NBA teams was extracted as a data frame from the NBA API. Important variables include: **Round Pick, Round Number, Overall Pick, College.**	4.6 MB	Pandas DF	API Documentation
College Stats	College statistics were extracted for each player from their player profile page on Sports Reference. These webpages were stored locally and parsed to isolate the Players Total table which contains: **Games (Played & Started), Field Goals (2-point & 3-point), Free Throws, Offensive & Defensive Rebounds, Assists, Steals, Blocks, Turnovers and Points, College Season, School and Conference.**	12.4 MB	Pandas DF	Sports Reference- College Basketball

DATA MANIPULATION METHODS

Clean and Prepare NBA Data for Regression Analysis Notebook Link

Data Retrieval & Filtering

We manually exported data from an online basketball stats repository called "Basketball Reference" to get contract values, this file included the yearly salary for each player through the year 2029. This analysis focused on current contract values, so we filtered the dataset for 2023-2024. We pulled player IDs and career stats from the NBA_API and filtered the datasets for current players with contracts.

Data Cleaning & Processing

Contract Value Dataset

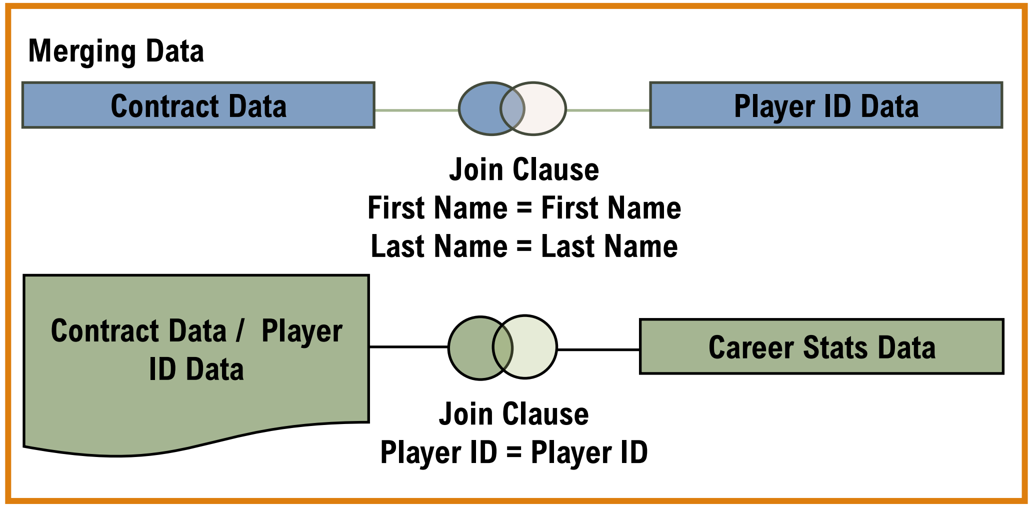
- Convert player names to lowercase characters.
- Replace accented characters with their English versions.
- Split the player column into first and last name columns.
- Rename columns to be more representative.
- Drop duplicate rows.

Player ID Dataset

- Convert first and last name columns to lowercase.
- Convert the player ID column to string datatype.
- Update specific player names to allow for joining.

Player Career Stat Dataset

- Convert the season column to datetime truncated to year.
- Drop rows corresponding to the 2024 season.
- Limit career stats to the last five years



Merging Data

Contract Data

Player ID Data

Join Clause
First Name = First Name
Last Name = Last Name

Contract Data / Player ID Data

Career Stats Data

Join Clause
Player ID = Player ID

Barriers to Data Manipulation

- Our dataset did contain null values because one player did not have a contract value, and several players are rookies competing in their first season this year; these players were dropped from the dataset because they did not fit the scope of our data goals.
- Several players needed name corrections because our data sources recorded names differently. The basketball reference website stores a player's name with accented characters, while the NBA_API stores the name without the accents. This was the most challenging data quality issue because it prevented the accurate joining of the two datasets. We addressed this using code that corrected specific names.
- The default file type for the player contract data is .xls, and this generates a value error because pandas cannot determine the file format. In this case, we opened the downloaded file, renamed it, and saved it as a .xlsx file

DATA MANIPULATION METHODS

Clean and Prepare College Data for College Stat Regression Analysis Notebook Link

Data Retrieval & Filtering

The goal of the associated notebook was to combine draft history variables with college statistics for each active player with an identified contract value who was drafted from a college/university. A pre-aggregated dataset was used as the primary source for players. To minimize the number of requests to the Sports Reference server, data needed to be retrieved in a particular sequence:

1. Extract team IDs from NBA API
2. Using team IDs, extract draft history for each NBA team and combine into one df
3. Keep only players in source list
4. The Draft History Endpoint imports all draft picks for each NBA team. Filter draft history to include only players who were drafted from a college/university.
5. Modify name values to match Sports Reference HTML links:
   - Remove spaces between last name and suffix
   - Remove apostrophes
   - Manual replacement of name values that don't follow typical naming format
6. Remove players without college record (attended but did not play or drafted from international college)
7. Scrape player profiles from Sports Reference

Once the name values were consistent with HTMLs, we extracted the entire player profile from Sports Reference for each player:

- For each player, pull HTML file and store locally (to avoid re-sending requests)
- Open each HTML file, parse and extract the "Players Totals" table into a df for each player
- Concatenate all player dfs together

Data Cleaning & Processing

The Players Total table contains a row for each college season and a totals row (to represent total stat values across all seasons). The provided totals row is dropped for the dataset. To mimic the aggregation of NBA stats, the season totals were summed by player – each player represented by exactly one row. Additionally, we engineered two additional variables from college stats:

- "Season Count": Count of college seasons
- "Team Count": Count of unique college teams

Merging Data

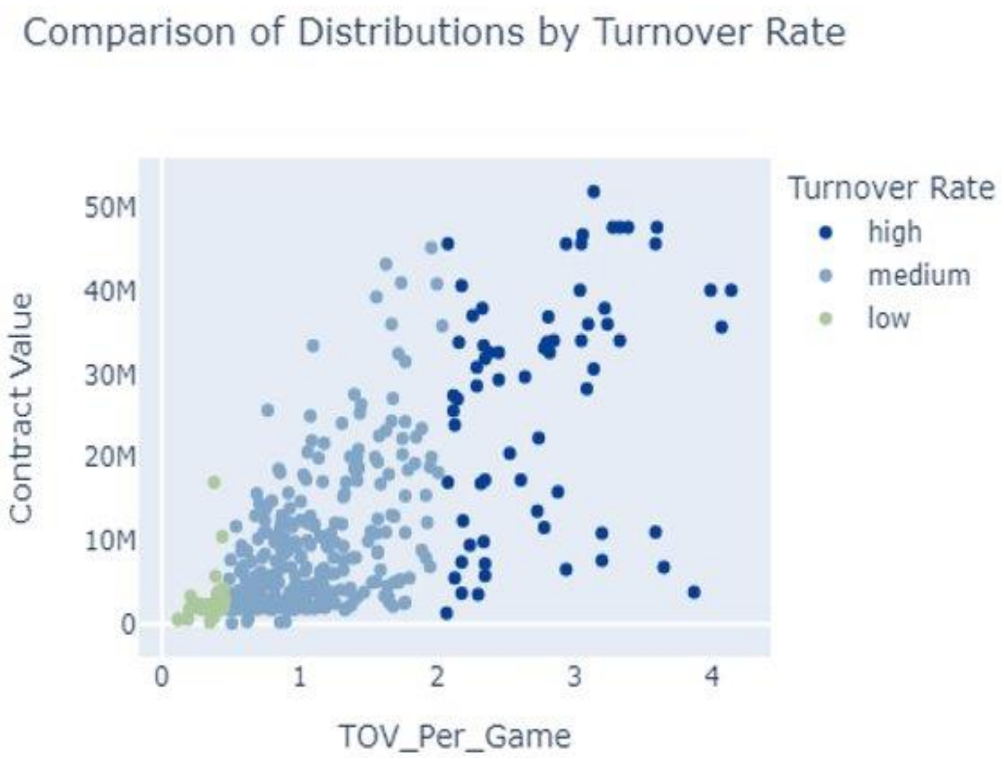
Finally, the college stats df was joined with draft history variables and contract values on player name.

Barriers to Data Manipulation

- The unique identifiers across dfs are not consistent (NBA API data contain unique player IDs while Sport Reference data are limited to players' names.
- Checking player name values against HTMLs was initially a very manual process – each time a request would fail, we would have to search the players name on the website and replace the value accordingly. Then we needed to re-run the request using the correct name value and manually delete the invalid HTML file.
- Even with proper HTMLs, the requests to the website server were extremely time-consuming – with execution times between 20 – 30 minutes per 15 players.

EXPLORATORY DATA ANALYSIS

Descriptive Statistics and First Insights Notebook Link - NBA



Interesting Insights and relationships
A moderately strong positive correlation exists between turnovers and a player's contract value. However, this case of correlation ***does not*** equal causation because turnovers hurt the game. When a player turns the ball over, his team loses an opportunity to score points. In the context of this dataset, players with higher contract values also have higher turnover rates because they usually have the ball more. It is ***not*** recommended that players increase their turnover rates.

The NBA regular season consists of 82 games, and the dataset contains metrics from the last five years, not including 2024, so the maximum number of games one can play is 410. The median games played across the 419 players in our dataset is just 206 games or 2.5 seasons. Players making more than $10 million per year tend to play more than 250 games over a 5-year period.

Steals per game have a weak positive correlation to contract values. Steals are obtained when a player assumes a defensive position and prevents the opposing player from scoring by taking the ball without fouling. Since this metric is associated with preventing the other team from scoring, it is surprising that players who steal the ball more do not have higher contract values.

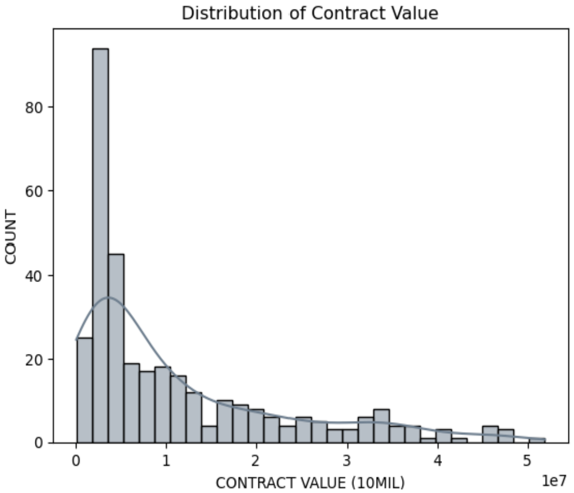
What didn’t work, and why?
In similar studies, additional metrics such as age, position, and years of experience are used to conduct the analysis. Unfortunately, I needed to timebox the time I spent looking for data sources and did not find one with the above stats. Additionally, I could not experiment with different regression algorithms for similar timebox constraints. Given more time, I would assess advanced NBA metrics and try different regression algorithms.

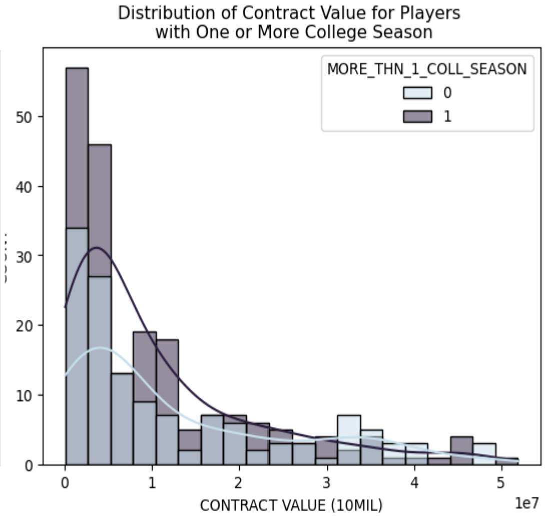
EXPLORATORY DATA ANALYSIS

Descriptive Statistics and First Insights Notebook Link - College

Wide Range of Contract Values

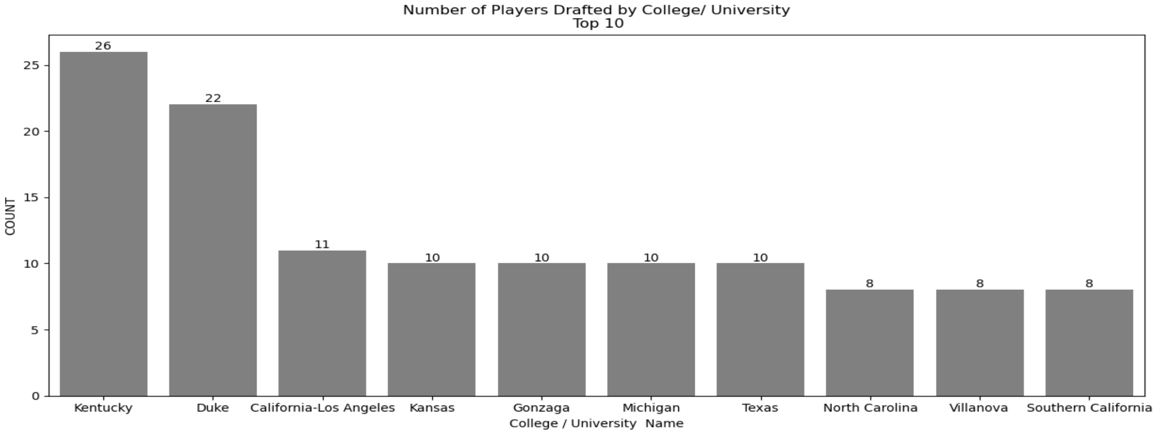
Not surprisingly, the range of contract values is quite large, with a minimum value of $92,857 and a maximum value of $51,915,615. This distribution is illustrated in the figure below. Most players in the NBA make below 10 million while a select few make 20 million or more. This distribution is exaggerated for players who played in more than one college season – which is most players. About 60% of players who played in college played more than one season. Even fewer of those players receive top-value contracts. This suggests an inverse relationship between contract value and number of college seasons played.





NBA Feeder Schools

There are 97 unique colleges/universities represented in our data. There was not any school that fed more than 10% of active players to the NBA. This was unexpected. In fact, the top three schools Kentucky, Duke and UCLA fed only (7.7%, 6.5 % and 3.2% respectively). The top ten NBA feeder schools are shown in the chart below.



Combination of Skills

The best-performing players are not always the highest earners. There are several instances of higher-performing players with relatively low value contracts within each stat. This suggest that a combination of skills may contribute more to a player's contract value that any one skill. Additionally, players stronger in offensive metrics (points and assists) tend to have higher contract values compared to players with high stats in steals or blocks.

REGRESSION ANALYSIS

Inferential Insights Notebook Link - NBA

NBA Linear Regression Process

1. Collect and pre-process data.
2. Conduct EDA to identify underlying trends.
3. Validate that the dataset meets the assumptions for regression analysis.
   - ○ Independent variables are linear to dependent variable.
   - ○ There is no multi-collinearity between independent variables.
   - ○ The observations in the dataset are independent of each other.
4. Scale independent variables.
5. Fit the model.
6. Remove insignificant features.
7. Fit the final model.

Final Model and Performance

We removed three points made, offensive rebounds, defensive rebounds, and steals per game because they were statistically insignificant. The final model uses field goals, assists, blocks, and personal fouls per game to explain the variation in contract values. Next, we used R-square, which measures how well the independent variables explain the variability of the dependent variable in a regression model to assess how well the final model explains contract values. **This model has an R-square of 70% which indicates only 30% of the variability in contract values is left to explain.**

Factors Impacting Contract Values

Standardize Feature	Interpretation
FGM_Per_Game: Field goals per game	Increasing average field goals by 1 **adds nearly 9 million dollars to contract value** when controlling for other factors.
AST_Per_Game: Assist per game	Increasing average assists by 1 **adds 1.3 million dollars to contract value** when controlling for other factors.
BLK_Per_Game: Blocks per game	Increasing average blocks by 1 **adds 1.2 million dollars to contract value** when controlling for other factors.
PF_Per_Game: Personal Fouls per game	Increasing average fouls by 1 **subtracts 9.8 million dollars from contract value** when controlling for other factors.

REGRESSION ANALYSIS

Inferential Insights Notebook Link - College

College Linear Regression Process

1. Collect and pre-process data
   - Feature engineering
   - Calculate per game level stats to replace college total stats
   - Calculate success rates for applicable metrics (Field Goals, Free Throws)
2. Remove correlated independent variables
3. Create dummy variables for School/University
4. Apply Standard Scaler to continuous independent variables
5. Fit Linear Regression model on scaled data

Final Model & Performance

As is, this model explains only 40% of the variability in contract value. We'd recommend additional model tuning before acting on any variables presented here. Of the 112 initial independent variables included in the linear regression model, only four were statistically significant: Overall Pick, Games Started, Points, and New Mexico State College. We expected the overall pick to be impactful, and it indicates market value for a player finishing their college career. Players picked later in the draft are typically less valued by NBA teams. Games Started, however, is surprising. More skilled players start more games and may expect higher contract values. We'd recommend a larger dataset and additional grouping for future model specifications to refine categorical variables.

Factors Impacting Contract Values

Standardize Feature	Interpretation
Overall Pick	Increasing a player's overall draft pick by 1 **decreases the contract value by nearly 3.5 million** when controlling for other factors.
Games Started	Increasing a player's average number of games started by 1 **decreases the contract value by nearly 2 million** when controlling for other factors.
Points	Increasing average pointes by 1 **adds 2.8 million dollars to contract value** when controlling for other factors.
New Mexico State College	If a player is drafted from New Mexico State, on average their **contract value increases by 22 million**

REFERNCES & STATEMENT OF WORK

References

- Vorkunov,M. (2023). Vorkunov: NBA salaries keep going up. Prepare to have your mind blown in the future. *The Athletic*. https://theathletic.com/4740069/2023/08/03/nba-salary-cap-rise-jaylen-brown/
- Feng, X., Wang, Y., & Xiong, T. (2023). NBA Player Salary Analysis based on Multivariate Regression Analysis. *Highlights in Science, Engineering and Technology*, *49*, 157-166. https://doi.org/10.54097/hset.v49i.8498
- Ando, K. (2018). NBA Players' Salary Prediction Using Linear Regression Model. https

Statement of Work

Damond Allen	Adrianna High
Pull NBA data from NBA_API Cleaning, manipulation and joining NBA data Visualization NBA data exploratory analysis Regression analysis on NBA data Setup Git Repo	Pull college data from NBA_API Cleaning, manipulation and joining college data Visualization College data exploratory analysis Regression analysis on college data Format final SlideDoc