Driving Success in the NBA: What Makes Winning Teams

*Farhad Ahmady (30194229), Rajvir Kaur (*30265204*), Marcelino Rodriguez Anglada (30265276), Matias Totz (30120305), Nyadual Makuach (30257163)*

(Photo by Chris Schwegler/NBAE via Getty Images)

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

The 2024 NBA draft where NBA teams select young basketball prospects to play on their teams, averaged 4.41 million viewers which was a record high. The average age of brand-new NBA players selected in the draft was just under 21 years old with the last drafted player earning $2.5 million dollars and the very first pick earning $12.6 million. For context, according to Statistics Canada, the median age of university graduates is just under 25 years old, and their average starting salary is $68,342 CAD. Unfortunately, there are only about 600 spots to play basketball in the NBA and the average player has a career of only 8.2 years. For the millions of people who aspire to make it into the NBA but don't make a spot, there are still a few options. In Canada, there is the Canadian Elite Basketball League which has roughly 120 spots open for basketball superstars with an average salary of $3,000 dollars per month. The only catch is that the league only runs for four months, and most people have no clue it even exists. However, since 2019 when the CEBL opened, 22 players have made it into the NBA. There is also the NBA G League for players who are not strong enough to make it into the NBA but are given a chance to develop in this league to potentially be called up to the NBA. Roughly 200 players are signed each year making on average $38,621 USD. For context, the average salary for a McDonald's crew member in the United States is around $28,317 USD per year and you can start working as early as 15 years old. You may be wondering what makes NBA players so special where they can earn more than $55 million dollars and over $100 million in endorsements like Stephen Curry of the Golden State Warriors. What is the difference between Tyler Smith who makes $1 million dollars and LeBron James who makes $48 million dollars. Why is LeBron James the same as 48 Tyler Smiths.

The 2024 NBA finals were watched by an average of over 11 million American viewers making it one of the most popular sports in the world. Basketball being so popular has had countless other studies investigate what determines salaries in the NBA. A study done by the Sport Journal found that the two most contributing factors to player salaries in the 2013-2014 NBA season were points per game and field goal percentage. This means that the most important factor in the NBA is not only scoring many points per game but also shooting on a high percentage (scoring 60% of the shots you take is better than scoring only 30% of the shots you take). In addition, getting rebounds, assisting other players, and personal fouls were statistically significant. The Journal of Sports Economics published a paper in 2010 which examined NBA players between the 1999-2000 and 2007-2008 seasons. Their study found that international players received a lower salary and player demographics such as height, positions, and draft order played a significant role in determining a player's salary. What is interesting is that both studies analyzed different aspects of the NBA over ten years ago. The first paper focused on game performance while the second paper focused on player demographics. Taking inspiration from both studies, our group wants to combine NBA players off court demographics with their on-court performance utilizing the most recent seasons in the NBA to create a model that best predicts a player's salary. We are interested in seeing if any new relationships form when you combine both off court demographics and on court performance using the most recent data rather than data that is over 10 years old.

1.1.2. Problem

The NBA is one of the most popular sports leagues in the world, with millions of fans tuning in to games and events every year. The 2024 NBA Finals alone had over 11 million American viewers on average, showing just how much people care about the league. The NBA also has a huge impact on the economy, from ticket sales and merchandise to the jobs created by arenas and local businesses that benefit from game days. NBA players themselves play a big role off the court too. For example, LeBron James donated millions of dollars to open the I PROMISE School, giving back to his community and inspiring others to do the same. Players earning higher salaries means more money can flow into charities and programs that make a difference.

Basketball also brings excitement into people's lives. Watching NBA games creates unforgettable moments, whether it's buzzer-beater shots or watching underdog teams succeed. The league isn't just about sports since it's about the joy and community it builds. That's why studying this topic matters and why we want to understand how salaries are determined so that more aspiring basketball players can make a bigger difference.

Right now, many studies on NBA salaries are outdated, looking at seasons that happened more than a decade ago. But the league has changed a lot in the last 10 years, with the rise of advanced analytics, shifts in play style, and new

priorities for teams. Older studies also tend to focus only on performance data or off-court factors like demographics, but rarely both together. Our project aims to fill this gap by combining these two areas to better understand what drives salaries today. We hope to uncover new patterns and relationships that reflect how the NBA operates now, rather than how it worked years ago.

1.1.3. **Challenges**

There are many different factors that affect how much money a player makes in the NBA. Some of these are obvious, like points per game or shooting percentage, but there are also a lot of variables that are harder to see. For example, how much a player helps their team's culture or how valuable they are for endorsements isn't something you can measure easily. On top of that, some things that you might think are important might not actually matter as much when we look at the numbers, which can make building a good model tough.

Outliers are also a big issue because they can really throw off the results. Take Bronny James as an example. He didn't have great stats in college and honestly wasn't that good, but he still got drafted because his dad is LeBron James, one of the biggest names in basketball. For context, in the six games Bronny James played in the NBA, he scored on average 0.7 points and had an average field goal percentage of 16 percent. One of the worst players in the league averaged 6.6 points which is significantly better than Bronny. LeBron's influence makes Bronny valuable to teams in ways that have nothing to do with his performance, like bringing in media attention and ticket sales. Variables like this shows how sometimes variables outside of basketball can matter more than what happens on the court. It has not been confirmed, but some NBA analysts believe that LeBron James promised teams that he would play for them if they drafted Bronny James. Sure enough, when the Lakers signed Bronny James, they also signed LeBron James.

The media can also have a huge impact on how much a player gets paid. If the media likes a player, it can really help their value, but if they don't, it can hurt them a lot. DeMarcus Cousins is a good example. He was an amazing player, but because the media focused on his attitude and painted him as a problem, he didn't get as many chances or the kind of money his talent deserved. It's not always about how good you are but also about how people see you.

Team chemistry is another factor that can change everything. If players don't get along, it can make the whole team worse, even if everyone is talented. A good example of this is when Kyrie Irving was with the Boston Celtics. He got injured during the playoffs, and the team played better without him. The younger players stepped up, and the chemistry improved, which shows how much relationships and teamwork can matter. Even if you're a star player, bad team chemistry can hurt your value.

1.1.4. **Additional Analysis: Coach Salaries**

While player salaries get a lot of attention, the salaries of NBA coaches also play a big role in the league. Coaches are the ones who design plays, build team chemistry, and make adjustments during games, but just like with players, figuring out what determines their pay can be tricky. A coach's success isn't always easy to measure because it depends on so many things, like the players they have or even how much the organization is willing to spend on winning. On top of this, all coach salaries are confidential other than a handful of coaches who have had their salaries leaked online.

One example of a coach is Nick Nurse, who led the Toronto Raptors to their first NBA championship in 2019. His success skyrocketed his value, earning him one of the highest-paid contracts among coaches at the time. But not every coach has the same opportunity to work with great players or big budgets, which means some talented coaches may not earn what they deserve.

Media and public perception can also play a role in how much a coach gets paid. A coach who builds a reputation for being innovative or a "players' coach" might have better chances of landing bigger contracts, even if their actual win-loss record isn't the best. On the flip side, a coach who struggles with team relationships or faces criticism for their decisions might find it harder to get paid what they're worth.

Another factor is how well a coach can manage egos and team dynamics. A good example of this is the Golden State Warriors' Steve Kerr, who not only led his team to multiple championships but also managed to balance the big

personalities of star players like Stephen Curry, Klay Thompson, and Kevin Durant. His ability to keep the team united under pressure adds value that goes beyond the games themselves.

Just like with players, our analysis on coach salaries will look at both performance factors, like win percentages, and off-court factors, like education. By understanding what impacts how coaches are paid, we can see how their contributions are valued in today's NBA. This smaller analysis complements our main study on player salaries, showing that both roles are essential in shaping a team's success.

1.2. OBJECTIVES

1.2.1. Overview

The main goal of our project is to figure out what factors influence how much NBA players and coaches get paid. For players, we're looking at both their performance on the court, like points per game and shooting percentage, and their off-court demographics, like their draft position and age. For coaches, we want to understand how factors like win percentages, education, and experience in the NBA play a role in their salaries. By combining these different areas, we hope to build models that can predict salaries more accurately than past studies, which often only focused on one piece of the puzzle.

Our project matters because the NBA is more than just a sport. It's a huge part of the economy, culture, and communities around the world. By understanding how salaries are determined, we can also see how the league values its players and coaches, which can have an impact on things like fairness, contracts, and even how money flows back into society through areas like charities.

1.2.2. Goals & Research Questions

1. Build a model that predicts NBA player salaries using a combination of on-court performance metrics (like points per game and true shooting percentage) and off-court demographics (like draft position, age, and country of origin).
2. Develop a smaller-scale model to analyze the factors that influence NBA coach salaries, focusing on performance (e.g., win percentage), reputation (e.g., public perception), and background (e.g., prior experience or education).
3. Figure out what are the most significant predictors of NBA player salaries when combining performance stats and player demographics?
4. What factors contribute most to the salaries of NBA coaches? Is it their win-loss record, reputation, or other variables like prior coaching experience?

2. METHODOLOGY

2.1 Data

Our NBA player analysis involved three datasets that were from various sources. These datasets were merged to create a final dataset with 22 columns and 1,468 rows, including player demographics, performance stats, advanced metrics, and salary. Our NBA coach analysis used two different data sources combined for a total of 28 columns and 33 rows for each coach. Listed below are the datasets used:

- **NBA players: This** dataset was derived from Kaggle. The dataset contains 27 seasons from 1996 to 2023. It has 36 teams with 2551 unique players before data cleaning.
- **Players Salary:** The dataset was derived from hoopsnype. Various component of the dataset was also taken from NBA website and combined. It has 12 columns and 1488 rows.
- **NBA Teams stats:** This dataset was derived from Kaggle.
- **Coaches:** It was derived from the 2022-23 NBA Coaches dataset from the website Basketball Reference
- **Coach Salaries:** Were taken from Salary Swish and if a coach did not have a salary listed, it was taken from wikipedia for that specific coach.

Agency that Generated the Data:

Kaggle: Kaggle is a platform known for hosting datasets across various domains, including sports, with contributions from data enthusiasts and professionals. The datasets used in this project were compiled from publicly available NBA statistics and shared on Kaggle by contributors aiming to provide accessible resources for basketball analysis.

HoopsHype: HoopsHype is a trusted platform in the basketball community that focuses on player salaries, contract details, and related news. It compiles salary data from NBA contract disclosures, media reports, and other official announcements.

Basketball Reference: Basketball Reference is a well-known platform that provides comprehensive statistics and historical data for professional basketball.

SalarySwish: SalarySwish is a public platform that aggregates salary data for NBA players and coaches from various reliable sources, focusing on transparency in sports earnings.

Wikipedia: Wikipedia was used to fill in gaps for specific coach salaries. While this source is widely accessible, its data is user-generated and might require further validation

Conditions Under Which Data Was Collected:

The Kaggle datasets were compiled from official NBA sources like box scores, team rosters, and performance statistics across 27 seasons.

The HoopsHype dataset aggregates salary information based on publicly available contract data, news articles, and official NBA announcements.

Basketball Reference collects data systematically, updating it based on official game and league outcomes.

SalarySwish compiles salary data from official disclosures and media reports.

Wikipedia entries are maintained by users and updated regularly but rely on cited sources for accuracy.

Sampling of the Dataset:

Kaggle: The dataset is systematic, covering all NBA players across the 27 seasons between 1996 and 2023. There is no random sampling; the dataset represents the complete population of NBA players during the specified timeframe.

HoopsHype: The salary data includes every player with a signed contract for the relevant seasons. The sampling method ensures no player is omitted, making the dataset comprehensive for analyzing salary trends.

The final dataset for players consists of 1,468 rows and 22 columns, representing every player and their performance metrics for the specified seasons. For coaches, the dataset includes 33 rows and 28 columns, representing every head coach for the 2022-23 season.

- For players: Each row represents an individual player for a specific season.
- For coaches: Each row represents an individual head coach for the 2022-23 NBA season.

Both our datasets for NBA players and NBA coaches used every single player and coach for the relevant seasons. Our final dataset from the combined datasets **player_salaries** consists of 22 columns and 1,468 rows and has the following attributes.

Player Demographics

- team_abbreviation: Abbreviation of the team the player belongs to (e.g., "TOR" for Toronto Raptors).
- age: Age of the player (in years).
- player_height: Height of the player (in centimeters).
- player_weight: Weight of the player (in kilograms).
- college: City where the college or university attended by the player is located.
- country: Country of origin of the player.
- draft_year: Year the player was drafted into the NBA, or "Undrafted" if not selected.
- draft_round: Round in which the player was drafted (e.g., "1" for the first round) or "Undrafted" if not selected.
- draft_number: Overall pick number in the draft (e.g., "8" for the 8th pick) or "Undrafted" if not selected.
- season: NBA season (e.g., "2020-21").

Performance Stats

- gp: Number of games played during the season.
- pts: Average points scored per game (points per game).
- reb: Average rebounds per game (includes both offensive and defensive rebounds).
- ast: Average assists per games,

Advanced Metrics

- net_rating: Net rating — the difference between the team's offensive and defensive ratings while the player is on the court.
- oreb_pct: Offensive rebound percentage — the proportion of available offensive rebounds secured by the player.
- dreb_pct: Defensive rebound percentage — the proportion of available defensive rebounds secured by the player.
- usg_pct: Usage percentage — an estimate of the percentage of team plays involving the player while on the court.
- ts_pct: True shooting percentage — a comprehensive measure of shooting efficiency accounting for field goals, three-pointers, and free throws.

Coach Data

- Education: The highest level of education completed by the coach or the institution attended.
- Salary: The total annual salary for the coach during the relevant season.
- Coach: Coach name
- Years as Coach: Year of experience as a coach in the NBA
- Career Win Rate: The total games won divided by the total games coached in the NBA
- Year: The season year
- Field of Study: What the coach studied in their highest education
- Played In NBA: If they played in the NBA when they were younger

Player Analysis:

- **Salary (Response Variable):** This is a quantitative and continuous variable representing the player's salary for a specific season in dollars. These directly impact how valuable a player is perceived by teams.
- **Age:** Measured in years, this is a quantitative variable.
- **Height and Weight:** Both are quantitative variables measured in centimeters and kilograms respectively.
- **Game Performance Stats (e.g., points per game, rebounds, assists):** Quantitative variables representing player performance on the court. These stats are continuous and directly influence salary.
- **Advanced Metrics (e.g., net rating, usage percentage, true shooting percentage):** Quantitative variables used to evaluate efficiency and contribution. These are calculated based on team and individual performance metrics. These provide deeper insights into player efficiency, especially for roles that don't solely focus on scoring.
- **Draft Details (round, number):** Categorical variables indicating a player's draft history, with "Undrafted" treated as a category.
- **Team Abbreviation and College:** Categorical variables indicating the team and educational background of the player.
-

Player Salary: This variable is quantitative and continuous. It is measured in dollars and does not represent a proportion, percentage, category, or binary value. It could theoretically take on decimal values, even if it doesn't in the dataset.

Coach Analysis:

- **Salary (Response Variable):** Quantitative and continuous variable representing the coach's salary in dollars.
- **Years as Coach:** Quantitative variable representing the number of years the coach has been coaching. These are key performance indicators, showing experience and success rates.
- **Career Win Rate:** A quantitative variable representing the proportion of games won during the coach's career.
- **Played in NBA:** Categorical variable indicating whether the coach has NBA playing experience. This can affect how players view the coach and may influence their ability to lead and strategize effectively.
- **Relevant Education:** Categorical variable indicating whether the coach has a background in relevant fields like Physical Education or Psychology. Indicates whether the coach has formal training in fields that might improve their coaching performance.

Coach Salary: Like player salary, this is quantitative and continuous, measured in dollars, and appropriate for regression modeling without requiring advanced techniques for binary or categorical data.

2.2 Approach

Player Analysis: We used multiple linear regression to model the relationship between the dependent variable, player salary, and several independent variables, such as player demographics, performance statistics, and advanced metrics. This approach allowed us to assess how each variable contributes to salary while controlling for others. By using coefficients, we could measure the direction and magnitude of these contributions, and statistical tests like p-values helped us identify significant predictors. Multiple linear regression was chosen because it is well-suited for analyzing continuous response variables and provides clear interpretability, making it consistent with the methods taught in DATA 603.

The order of our analysis started with data preparation, where we combined datasets, cleaned missing values, and filtered by season to analyze data for 2020-21, 2021-22, and 2022-23 separately. Next, we created a full model that included all possible predictors. Using backward elimination, we simplified the model by removing predictors with p-values greater than 0.05. This step ensured that only significant predictors were included in the reduced model. We then compared the reduced model to the full model using ANOVA tests to confirm that excluded predictors were not statistically significant. After obtaining the best additive model, we introduced interaction terms to check if combinations of variables (e.g., age and points per game) had a significant effect on salary. Finally, we tested higher-order terms, like

quadratic relationships, for key predictors to capture any non-linear effects. Each step depended on the results of the previous one, making the workflow logical and consistent.

We used several techniques to enhance our analysis and justify their application. Backward elimination helped us focus on significant predictors, making the model simpler and easier to interpret. ANOVA testing validated that removing certain predictors did not weaken the model. Interaction terms were used to explore how combined variables influenced salary, as some relationships might not be purely additive. Higher-order terms, such as quadratic transformations, were applied to identify non-linear patterns for predictors like points per game or games played. Residual diagnostics were performed to check assumptions like normality, independence, and homoscedasticity, ensuring the model met the requirements for linear regression. These methods provided a robust framework for our analysis, aligning with techniques covered in DATA 603.

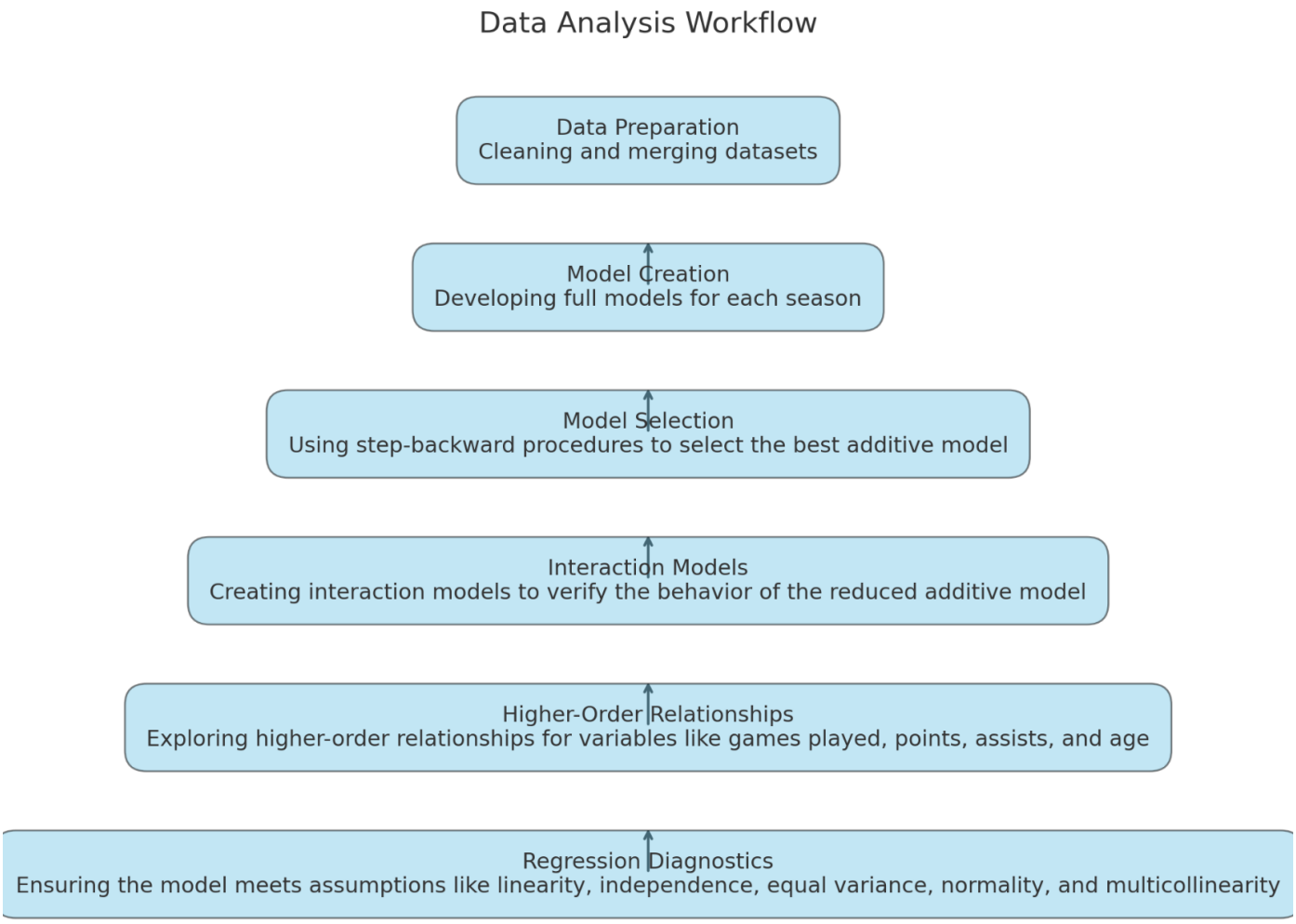
We used an alpha value of 0.05 for all hypothesis tests to determine statistical significance. This standard threshold allowed us to decide whether predictors were significant contributors to the response variable, ensuring consistency across all steps of the analysis. Predictors with p-values below 0.05 were retained, while those above this threshold were removed during backward elimination. This systematic approach ensured the model was both statistically valid and easy to interpret.

Coach Analysis: First, we need to load the data and filter for 2022-2023. Originally, we were going to look at two different years, but we ran into problems because there were coaches who were the coach of a team for both years. This was throwing off our results, so we decided to go with 2022-2023 since it had the most complete coach salaries. Next, we decided to define what were relevant education paths for coaches in the NBA so that we could use it for further analysis. We then combined the original data with the salary data for coaches. We then need to build a full additive model with all relevant variables. Then we will look at an interaction model, summarize both models and clean out any non-statistical variables. Then we can run an anova test to see which model is better. Then we can summarize the adjusted r square and RSE of each model and compare them. This will then lead us to the final model.

The biggest issue we ran into during our Data 604 project was that none of our results for coaches were statistically significant. This may cause major issues in this project so we will ensure to include interactive models to get as close as possible to a complete model. We also are just using the coach analysis as a secondary analysis and will not spend nearly as much time as we would on the player analysis. This may impact how far we take our final model, and we may end up with a model that only accurately predicts 30-50% of a coach's salary accurately.

2.3 Workflow

Below illustrates the steps we followed for data analysis process during our project exploration.



Techniques Used and Justification:

- **Multiple Linear Regression**: Quantifies the effect of predictors on salary.
- **Stepwise Regression**: Efficiently identifies significant predictors by removing non-significant ones.
- **ANOVA Tests**: Validates model simplifications by comparing reduced and full models.
- **Interaction Terms**: Test whether the effect of one predictor depends on another.
- **Polynomial Terms**: Check for non-linear relationships, although many were not significant.

Alpha Value: A significance level of 0.05 is used consistently in hypothesis testing.

As stated earlier, the analysis of coach salaries was just an extra piece added to the project. Due to time constraints, we did not nearly go as far as we did with the player analysis. We mostly used the same flow chart however we ended up at the interaction models as our final step and did not use any backward step functions.

2.4 Contributions

Briefly describe the group members' workload distribution and responsibilities.

Farhad Ahmady: Farhad will focus on the regression diagnostics for our player analysis and will finish 4.1 and 4.2 of our final report. He will also oversee double checking 3.1 and ensuring each step has been covered. Farhad will also help prepare and write a script for the final presentation based on the results of our analysis.

Rajvir Kaur and Marcelino Rodriguez Anglada: Rajvir and Marc will focus on our player analysis up until regression diagnostics. Both members have either a master's or undergraduate degree in software engineering and have experience writing code. They will also focus on writing the results in the 3.1 section of the final report.

Matias Totz: Matias will be in charge of writing 1.1 motivatation and 1.2 objectives for the final report. Since he has the best english skills, he will also be in charge of fixing all grammar mistakes in our final report draft. Matias has experience working in the consulting industry and will be in charge of creating final presentation slides. Matias will also lead the exploration of coach salary analysis.

Nyadual Makuach: Nyadual will be in charge of writing 2.1, 2.2, 2.3, and 2.4. She will also help with the analysis on coach salaries and create all the visuals for the final report. With the most experience in creating references, she will also oversee correctly referencing all sources and images in our final report and presentation. Nyadual will also oversee adding the correct comments into our final RMD file.

3 MAIN RESULTS OF THE ANALYSIS

3.1 Results

What do my results indicate? Do you have any unexpected results? Please elaborate.

In this analysis, we performed a Multiple linear regression where the dependent variable is salary, and the independent variables include age, player height, player weight, games played (GP), points (PTS), rebounds (REB), assists (AST), net rating, offensive rebound percentage (OREB_PCT), defensive rebound percentage (DREB_PCT), usage percentage (USG_PCT), true shooting percentage (TS_PCT), assist percentage (AST_PCT), team abbreviation, college, country, and draft number.

1. **Create a full additive model that contains both dependent and independent variables.**

The multilinear regression model is used to analyze the relationships between the variables in the dataset(independent) and the response variable, which in this case is Salary. Salary serves as the dependent variable, and the goal is to understand how it is influenced by the other variables. Initially, the full model includes all variables from the dataset. However, the data has been cleaned and filtered to focus on specific seasons. For this study, the analysis is centered on the 2020-2021 season.

```
Call:
lm(formula = salary ~ age + player_height + player_weight + gp +
    pts + reb + ast + net_rating + oreb_pct + dreb_pct + usg_pct +
    ts_pct + ast_pct + team_abbreviation + college + country +
    draft_number, data = players_salaries_Variable)
```

Figure 1 Full Model

Figure 1 shows the full created model in R studio, it shows all the possible predictors.

```
Residual standard error: 5.753 on 230 degrees of freedom
Multiple R-squared:  0.8078,    Adjusted R-squared:  0.6297
F-statistic: 4.537 on 213 and 230 DF,  p-value: < 2.2e-16
```

Figure 2 Full Model Results

Figure 2, shows the results for the full model:

- Adjusted R-squared (0.6297): Indicates that approximately 62.97% of the variation in the dependent variable (player salaries) can be explained by the model, accounting for the number of predictors.
- p-value (< 2.2e-16) is less than 0.05 at significance level, suggesting that the model is statistically significant.

Once the Models had been created. It is possible to use the selection methods to create and choose the best additive model for this analysis.

In this case, **Step Backward Procedure** is the one being used to select the best additive method:

```r
backward_model_2020_2021=ols_step_backward_p(full_model_2020_21, p_val
0.05)
```

Figure 3 Step Backward Procedure

Figure 3, shows the code used to run the Step Backward Procedure in R Studio.
Once the code has been executed, the summary is study to draw some conclusions. Figure 4, shows the results from the Step Backward Procedure.

```
Call:
lm(formula = paste(response, "~", paste(c(include, cterms), collapse = " +
")),
    data = l)

Residuals:
     Min       1Q   Median       3Q      Max
-18.8621  -3.2901  -0.0413   2.8946  20.4295

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -38.24184    6.98503  -5.475 7.42e-08 ***
age             0.63705    0.06332  10.061  < 2e-16 ***
player_height   0.12131    0.03393   3.575  0.00039 ***
gp             -0.05204    0.01571  -3.312  0.00100 **
pts             0.65533    0.06547  10.010  < 2e-16 ***
ast             1.40450    0.22205   6.325 6.30e-10 ***
draft_number2  -1.16745    0.77091  -1.514  0.13066
draft_number3  -2.06649    0.82939  -2.492  0.01309 *
draft_number4  -2.66862    0.97641  -2.733  0.00653 **
draft_number5  -3.52393    0.81123  -4.344 1.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.464 on 434 degrees of freedom
Multiple R-squared:  0.6729,    Adjusted R-squared:  0.6661
F-statistic: 99.18 on 9 and 434 DF,  p-value: < 2.2e-16
```

Figure 4 Results from Step Backward Procedure

The results obtained from Step Backward Procedure are:

- Adjusted R-squared (0.6661): This means that approximately 66.61% of the variation in the dependent variable (player salaries) is explained by the model, after adjusting for the number of predictors.
- p-value (< 2.2e-16): small p_value less than 0.05 at significant level suggests that the model is statistically significant. This means there is strong evidence to suggest that the predictors in the model are having a meaningful effect on the dependent variable.

From the results obtained from Step Backward Procedure, the original full model got reduced to a less complex model.

This new reduced model behaves better than the original model because it has a better R adjusted square and less predictors to analyze making it less complex. From figure 4 it is evident that almost all predictors are significant and the one that is not significant (draft_number2) is a categorical variable therefore it cannot be dropped out of the model. However, to verify that the already dropped predictors were accurately dropped it is necessary to create a hypothesis test.

Create Anonova table to test the new additive model:

The Annova table compares 2 models and analyzes the p value to make sure that the changes in the predictors are appropriate and do not impact the model negatively. To verify the mentioned the Annova table is being used to compare the full model and the additive model created by the Step Backward process.

```r
anova(reduce_model_2020_21, full_model_2020_21)
```

```
Analysis of Variance Table

Model 1: salary ~ age + player_height + gp + pts + ast + draft_number
Model 2: salary ~ age + player_height + player_weight + gp + pts + reb +
    ast + net_rating + oreb_pct + dreb_pct + usg_pct + ts_pct +
    ast_pct + team_abbreviation + college + country + draft_number
  Res.Df     RSS  Df Sum of Sq      F Pr(>F)
1    434 12955.0
2    230  7612.4 204    5342.6 0.7913 0.9564
```

Figure 5 Annova Table

Figure 5 shows the creation of the Annova table and its respective results. The results from the Annova table are being used to define a hypothesis test.

- The null hypothesis tests that all the dropped predictors are insignificant (equal to zero).
- The alternative hypothesis tests that at least one predictor is significant (different from zero) for this model.

$$H_0 : B_1 = B_{2...} = B_{droppedPredictors} = 0 \quad VS \quad H_a : B_i \neq 0$$

Figure 6 Hypothesis Statement for Annova Table

The results obtained from anonova table (Figure 6):

- We found the p-value is 0.9564 which is bigger than alpha at 0 .05. Then we fail to reject the null hypothesis and conclude that the dropped predictors were insignificant and accept our reduced model with Step Backward Procedure.
- The significant predictors obtained from the **Step Backward Procedure** are age, player_height, gp, pts, ast, and draft_number.

2. **Create an Interaction model on the significant predictors.**

The interaction model is a model that studies the predictors and its interactions, to create this model the square is being used. See Figure 7.

```r
interactive_reduce_model <- function(players_salaries_Variable) {

  intereactive_model <- lm(salary ~ (age + player_height + gp + pts + ast +
draft_number)^2, data=players_salaries_Variable )

}
```

Figure 7 Interaction Model

In the interaction model the variables or predictors that are being used are the ones that were selected from the Step Backward Procedure (model selection) and tested by the Annova table. Figure 8 shows the results from the interactive model.

```
Call:
lm(formula = salary ~ (age + player_height + gp + pts + ast +
    draft_number)^2, data = players_salaries_Variable)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5373  -2.2926  -0.1398   1.5112  21.3932

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.325e+01  4.338e+01  -0.306  0.76013
age                         6.183e-01  1.472e+00   0.420  0.67471
player_height               1.176e-01  2.112e-01   0.557  0.57788
gp                         -1.251e-01  3.713e-01  -0.337  0.73635
pts                        -3.605e+00  1.372e+00  -2.627  0.00894 **
ast                        -3.165e+00  4.303e+00  -0.736  0.46237
draft_number               -1.911e-01  4.269e+00  -0.045  0.96431
age:player_height          -4.361e-03  7.123e-03  -0.612  0.54073
age:gp                     -5.209e-03  3.034e-03  -1.717  0.08670 .
age:pts                     8.162e-02  1.432e-02   5.700 2.25e-08 ***
age:ast                     1.400e-01  4.244e-02   3.298  0.00106 **
age:draft_number            3.315e-02  4.082e-02   0.812  0.41720
player_height:gp            7.628e-04  1.834e-03   0.416  0.67763
player_height:pts           1.127e-02  6.299e-03   1.789  0.07435 .
player_height:ast           3.006e-03  2.027e-02   0.148  0.88219
player_height:draft_number -7.182e-03  2.069e-02  -0.347  0.72870
gp:pts                      2.471e-03  3.277e-03   0.754  0.45116
gp:ast                      1.013e-02  1.240e-02   0.817  0.41457
gp:draft_number             1.455e-02  9.347e-03   1.557  0.12033
pts:ast                    -2.331e-02  2.241e-02  -1.040  0.29884
pts:draft_number           -7.959e-02  3.969e-02  -2.005  0.04557 *
ast:draft_number           -1.094e-01  1.436e-01  -0.762  0.44668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.536 on 422 degrees of freedom
Multiple R-squared:  0.7807,    Adjusted R-squared:  0.7698
F-statistic: 71.54 on 21 and 422 DF,  p-value: < 2.2e-16
```

Figure 8 Interactive Model Results

The results obtained from the above model are Adjusted R-squared: 0.7698 and p-value:2.2e-16.

- Adjusted R-squared (0.7698): This value indicates that approximately 76.98% of the variation in the dependent variable is explained by the interactive model, after adjusting for the number of predictors.

- p-value (< 2.2e-16): The small p-value indicates that the model as a whole is statistically significant.

From the results in the interactive model there is one interesting condition that needs to be met.

- Some predictors became insignificant. However, due to hierarchical rules these father predictors must remain in the model.

Besides the father predictors, all other predictors that hold p-values are less than 0.05 need to be removed. To conclude the creation of the interactive model the predictors age, player_height, gp, pts, ast, draft_number, age*pts, age*ast, pts*draft_number are all significant predictors for the model. Therefore, a new Reduce Interactive Model is created after dropping the insignificant predictors. Figure 9 shows the new reduced interaction model.

```
Call:
lm(formula = salary ~ age + player_height + gp + pts + ast +
    factor(draft_number) + age * pts + age * ast + pts * draft_number,
    data = players_salaries_Variable)
```

Figure 9 Reduced Interaction Model

Same as the previous section it is necessary to test that all the predictors dropped were accurate drop and not impacting the model negatively. . To verify the mentioned the Annova table is being used to compare the two different interactive models. See figure 10.

```
```{r}
anova(reduce_interactive_model_2020_2021,interactive_reduce_model_2020_2021)
```

Analysis of Variance Table

Model 1: salary ~ age + player_height + gp + pts + ast +
factor(draft_number) +
    age * pts + age * ast + pts * draft_number
Model 2: salary ~ (age + player_height + gp + pts + ast + draft_number)^2
  Res.Df   RSS Df Sum of Sq   F Pr(>F)
1   431 9094.6
2   422 8684.6 9   410.02 2.2137 0.02035 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10 Annova Table for Interactive Models

The results from the Annova table are being used to define a hypothesis test (figure 11).

- The null hypothesis tests that all the dropped predictors are insignificant (equal to zero).
- The alternative hypothesis tests that at least one predictor is significant (different from zero) for this model.

$$H_0 : B_1 = B_{2...} = B_{droppedPredictors} = 0 \quad VS \quad H_a : B_i \neq 0$$

Figure 11 Annova Hypothesis Test Int_Mod

From the results from the Anova table we found that the p-value is 0.02035 which is less than alpha at 0.05. Then we reject the null hypothesis and conclude that the dropped predictors were insignificant and accept the reduce interaction model.

The significant predictors obtained from the Interaction model are age, player_height, gp, pts, ast, draft_number, age:pts, age:ast, pts:draft_number.

3. Use of Higher order terms

The next step is to verify if the model can have any of the terms a higher order. To do this the GGally package is used. To verify that a predictor could be relevant to apply higher level ggpairs command has been used. See figure 12.

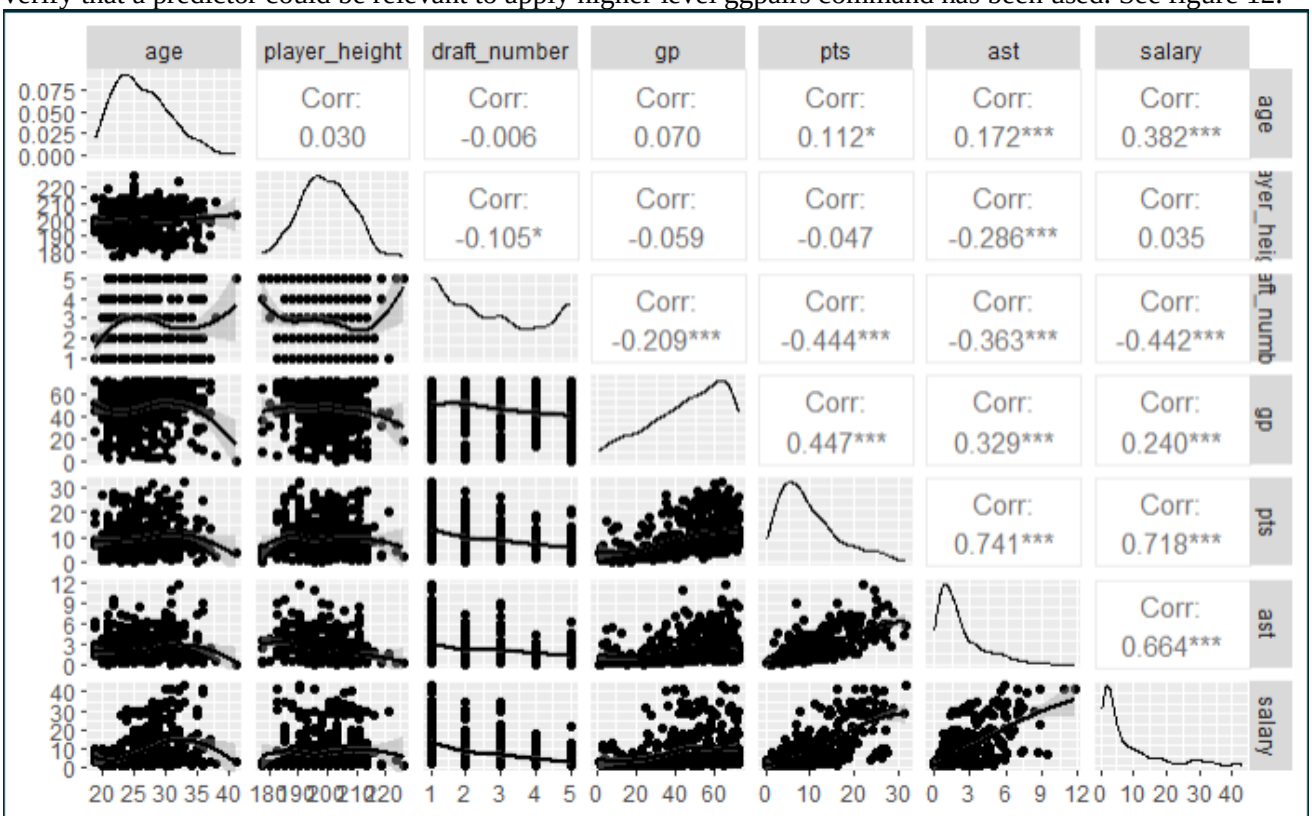


Figure 12 Independent Variable to Response Variable

Figure 12 shows how independent variables behave against the response variable, it is being used to detect patterns and parabolic behaviors. Behaviors like the ones mentioned could indicate a higher order relationship between the variables.

From figure 12 a clear pattern was not detected therefore the higher order relationship was applied to all the variables. However, only one variable (age) was successfully accepting quadratic value. See figure 13.

```
Call:
lm(formula = salary ~ age + player_height + gp + pts + I(age^2) +
    ast + draft_number + age * pts + age * ast + pts * draft_number,
    data = player_salaries_2020_21)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9316  -2.3541  -0.2491   1.7631  20.3936

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -36.27358    9.88424  -3.670 0.000273 ***
age                1.22954    0.58548   2.100 0.036301 *
player_height      0.12559    0.02813   4.464 1.03e-05 ***
gp                -0.03452    0.01355  -2.547 0.011211 *
pts               -0.90155    0.34390  -2.622 0.009061 **
I(age^2)          -0.02756    0.01040  -2.649 0.008373 **
ast               -2.92549    1.04415  -2.802 0.005310 **
draft_number      -0.36985    0.26441  -1.399 0.162592
age:pts            0.06590    0.01244   5.298 1.87e-07 ***
age:ast            0.15027    0.03710   4.051 6.05e-05 ***
pts:draft_number  -0.07344    0.02724  -2.696 0.007287 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.55 on 433 degrees of freedom
Multiple R-squared:  0.7736,    Adjusted R-squared:  0.7684
F-statistic:   148 on 10 and 433 DF,  p-value: < 2.2e-16
```

Figure 13 Higher Order Relationship Quadratic Age

Figure 13 shows the predictors used in the interaction model and shows the new Higher Order value (age^2). After applying this change to the model, it is important to keep a record of the new values shown in figure 13.

The results obtained from the above model are Adjusted R-squared: 0.7684 and p-value:2.2e-16.

- Adjusted R-squared (0.7684): This value indicates that approximately 76.84% of the variation in the dependent variable is explained by the interactive model, after adjusting for the number of predictors.
- p-value (< 2.2e-16): The small p-value indicates that the model is statistically significant.

4. Regression Diagnostics

In this section assumptions are investigated to confirm that the selected model meets all the assumptions and is well justified.

4.1 Linear Assumption

The linear regression model assumes that there is a straight-line (linear) relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect and the prediction accuracy of the model can be significantly reduced.

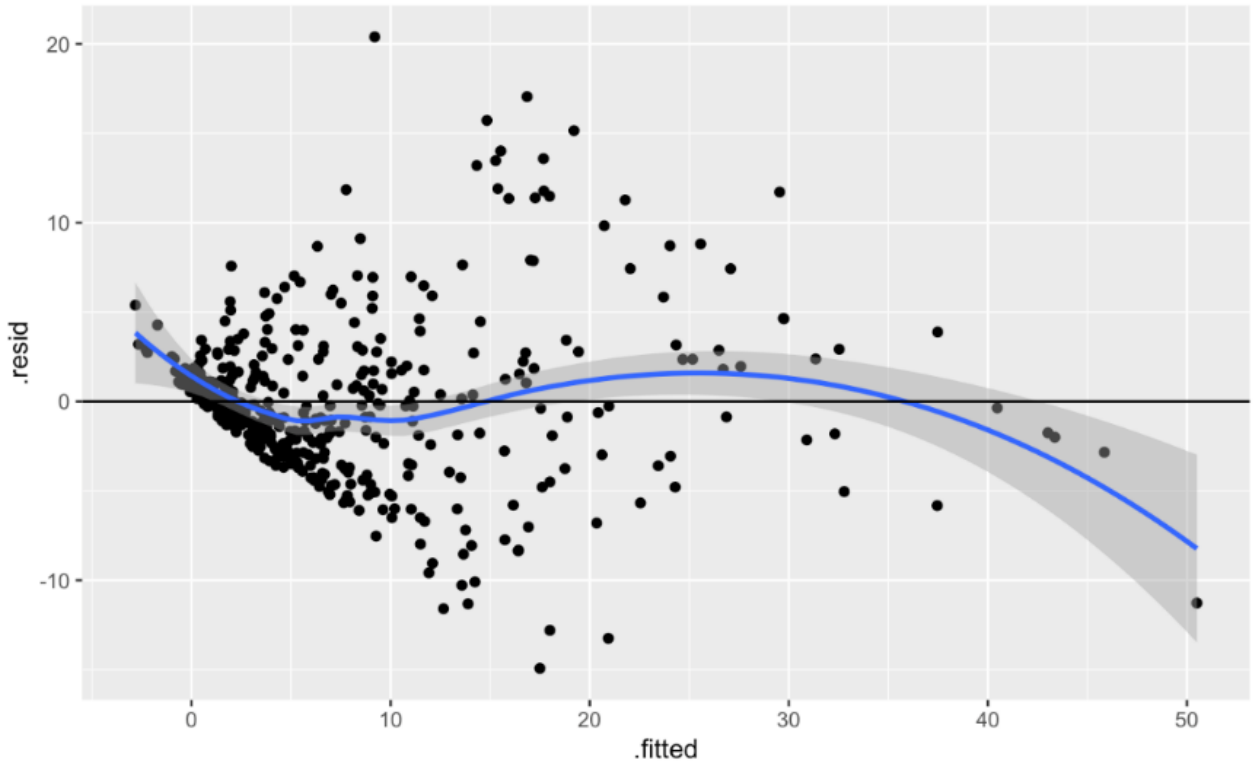


Figure 14 Linear Regression Line

Figure 14 shows the regression line and from its behavior and because the higher order relationship has been evaluate, it is possible to conclude that the model meets the linear assumption.
The adjuster r squared for the quadratic model for age is 0.7684 indicates the variation in salary that can be explained by this model is 76% with RMSE 4.55.

4.2 Independence Assumption

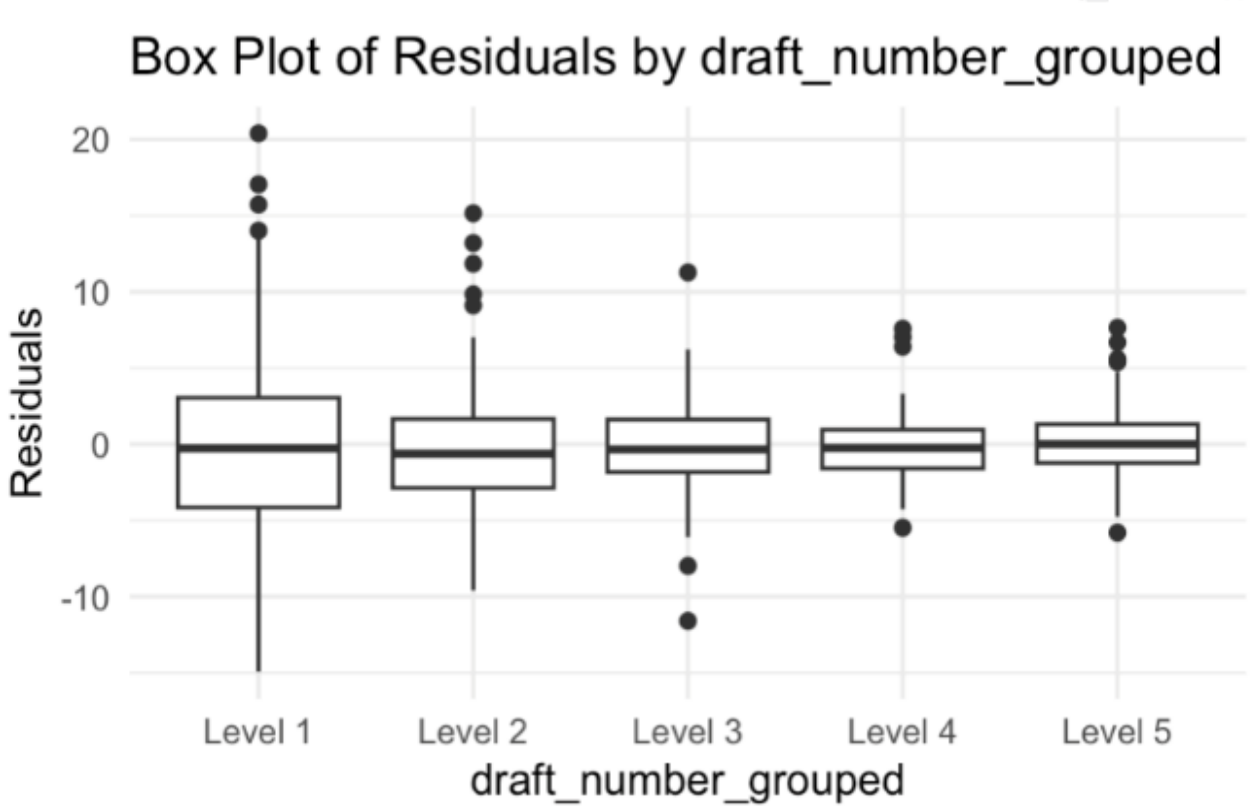


Figure 15 Drop Box Independence Assumption

Figure 15 shows how the data behaves and how independent it is. In this case we have a categorical variable that groups values, this has been used to reduce the frequency of the data and create a more global extended group to better understand and analyze the data. After grouping and defining the categorical variable from figure 15 it is possible to conclude that the model meets the independence assumption.

4.3 Equal Variance Assumption

Equal variance test that the data behaves the same over the range of measured values. There are multiples option to verify this assumption. One of the methods to verify homoscedasticity is using graphs, However using this can be hard to read. The figure 16 shows the graph for this assumption:

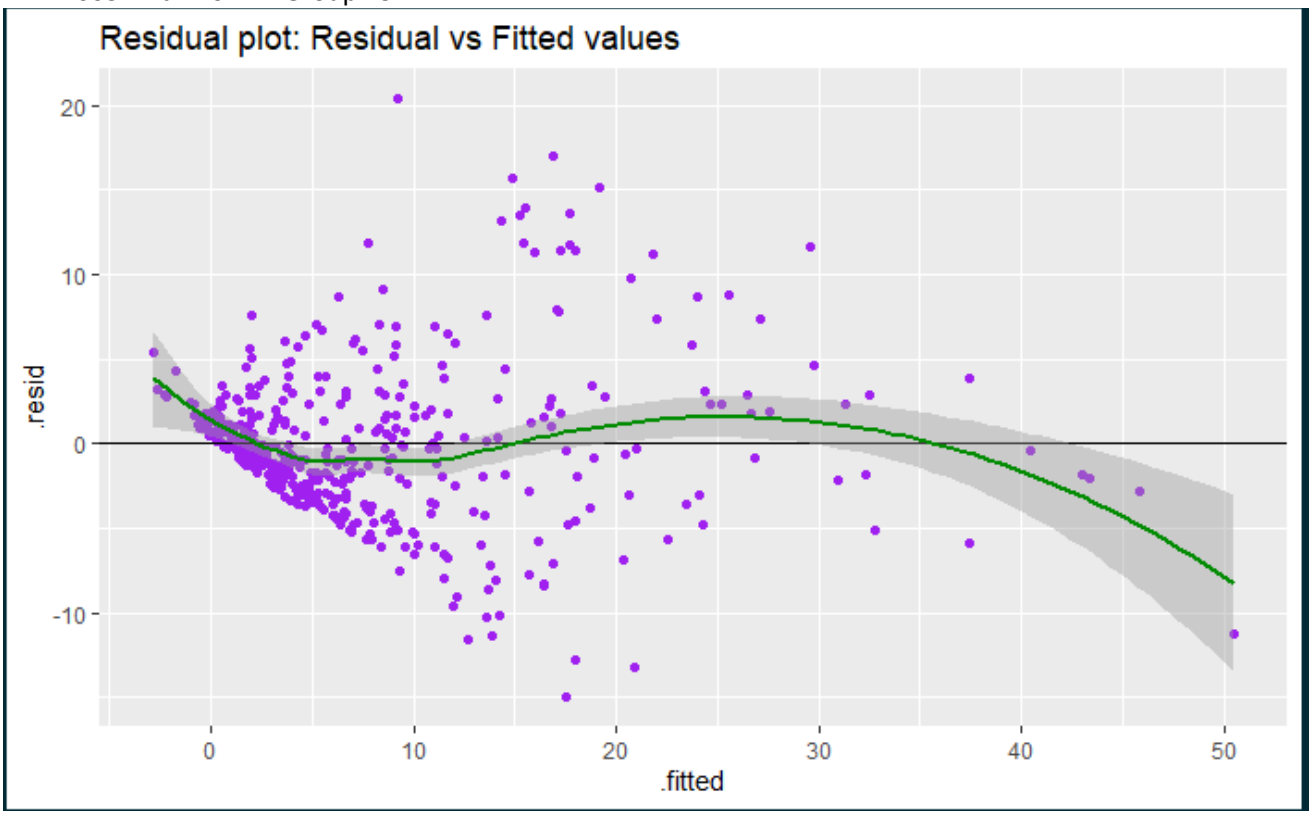


Figure 16 Homoscedasticity Assumption

From the graph it seems like the data is changing therefore it might have homoscedasticity. To verify what the graphs show another method is being. The Breusch-Pagan Test is used to verify heteroscedasticity.

```
```{r}
bptest(age_higer_model_2020_2021)
```

        studentized Breusch-Pagan test

data: age_higer_model_2020_2021
BP = 78.117, df = 10, p-value = 1.173e-12
```

Figure 17 The Breusch-Pagan Test

From figure 17, it is possible to see the result, to understand this result it needs to be analyzed against a hypothesis test.

$$H_0 : \sigma\frac{2}{1} = \frac{2}{n} \quad VS \quad H_a : \sigma\frac{2}{i}; \quad where \quad i = 1, 2...n$$

Figure 18 Hypothesis Test for Breusch-Pagan Test

This number is smaller alpha at 0.05, testing the hypothesis:

- Null hypothesis: heteroscedasticity is not present (homoscedasticity)

- Alternative hypothesis: heteroscedasticity is present

From figure 17, the result is 0.02876 this is smaller than alpha at 0.05, therefore we reject the null hypothesis and conclude that homoscedasticity is present.

There are some methods that can be used to solve the homoscedasticity presence. In this case the method selected is applying the Log Transformation.

```
```{r}
bcmodel_log_reduce=lm(log(salary) ~ age + player_height + gp + pts +
I(age^2) + ast + draft_number + pts*draft_number,
data=player_salaries_2020_21)
summary(bcmodel_log_reduce)
```


Call:
lm(formula = log(salary) ~ age + player_height + gp + pts + I(age^2) +
    ast + draft_number + pts * draft_number, data =
player_salaries_2020_21)
```

Figure 19 Log Transformation

Figure 19 shows the log transformation being applied to the method. With this new created method apply the Breusch-Pagan Test again to verify that the homoscedasticity is not present.

```
```{r}
bptest(bcmodel_log_reduce)
```


        studentized Breusch-Pagan test

data: bcmodel_log_reduce
BP = 17.133, df = 8, p-value = 0.02876
```

Figure 20 Breusch-Pagan Test to Log Model

Figure 20 shows the application of the Breusch-Pagan Test to the log model, showing a bigger p value that the previous model, however the p value is still small than alpha at 0.05. Therefore, we reject the null hypothesis and conclude that homoscedasticity is still present. In conclusion the analyze model does not meet the Equal Variance Assumption and it needs deeper study to solve the issue.

4.1 Normality Assumption

This assumption check for the data to be normally distributed, this can be check using graphs and the Shapiro-Wilk test (S-W).
It is important to keep in mind that the log transformation was applied to the model, that model is the one that it is being used form know on in the analysis.

```r
ggplot(data=player_salaries_2020_21, aes(residuals(bcmodel_log_reduce))) +
geom_histogram(breaks = seq(-1,1,by=0.1), col="green3", fill="green4") +
labs(title="Histogram for residuals") +
labs(x="residuals", y="Count")
```

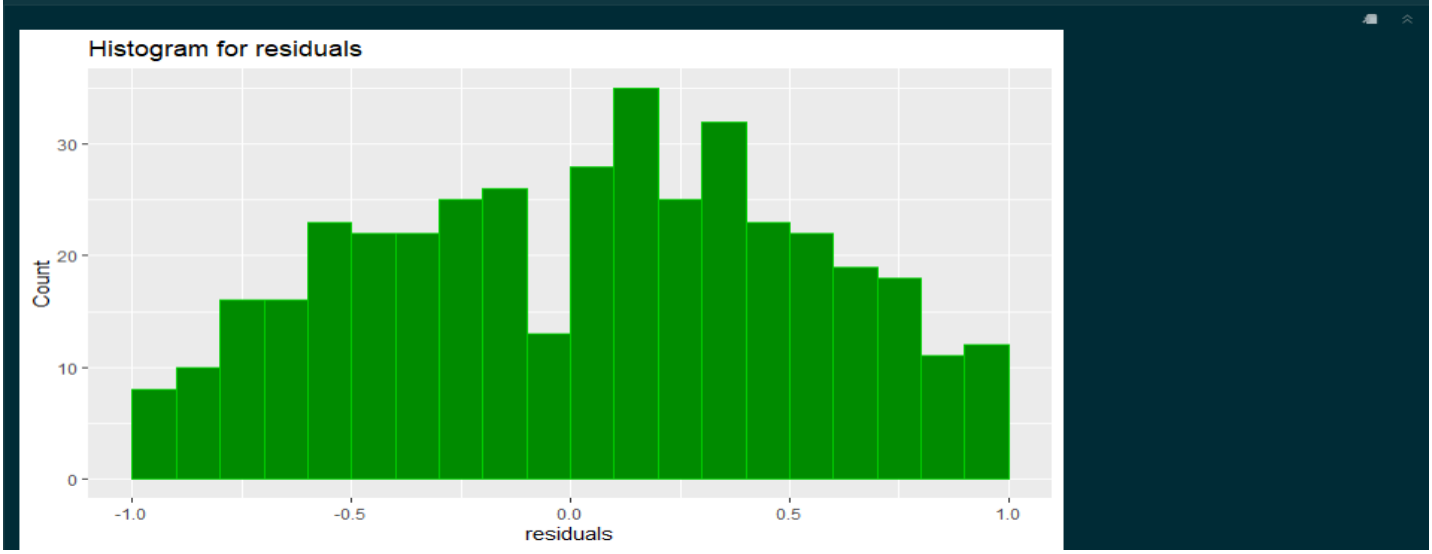


Figure 21 Histogram for Normality Assumption

From figure 21, it is possible to deduce that the data follows the Normality assumption. However, it is important to run the Shapiro-Wilk test to verify:

```r
shapiro.test(residuals((bcmodel_log_reduce))
```

```
        Shapiro-Wilk normality test

data:  residuals((bcmodel_log_reduce))
W = 0.994, p-value = 0.07742
```

Figure 22 Shapiro-Wilk test (S-W)

Figure 22 provides the results of the Shapiro test, then we test the results:

- Null hypothesis: H_0: The sample data are significantly normally distributed.
- Alternative hypothesis: H_1: The sample data are not significantly normally distributed

From figure 22, it is being observed that a P value of 0.07742 is bigger than the alpha at 0.05 therefore, we fail to reject the null hypothesis. In other words, the data is significantly Normal distributed. In conclusion, our model meets the Normality Assumption

4.2 Multicollinearity Assumption

During this analysis, it is expected to have a Multicollinearity, due to the interactions between variables as I(age^2) that would be strongly related to the age predictor. However, this interaction is being ignored if is shown in the analysis.

```r
imcdiag(bcmodel_log_reduce, method="VIF")
```

```
Call:
imcdiag(mod = bcmodel_log_reduce, method = "VIF")


 VIF Multicollinearity Diagnostics

                      VIF detection
age                124.2594         1
player_height        1.2187         0
gp                   1.3576         0
pts                  5.4319         0
I(age^2)           123.8949         1
ast                  2.7178         0
draft_number         3.5577         0
pts:draft_number     3.8737         0

Multicollinearity may be due to age I(age^2) regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Figure 23 Multicollinearity Diagnostics

Figure 23 shows that the Multicollinearity between the variables age and $I(age^2) $ is present. This is to be ignore, that being said the model meets the Multicollinearity Assumption.

4.3 Outliers

In this section analyze the outliers if present and if they are influential in the behavior of the data and the model

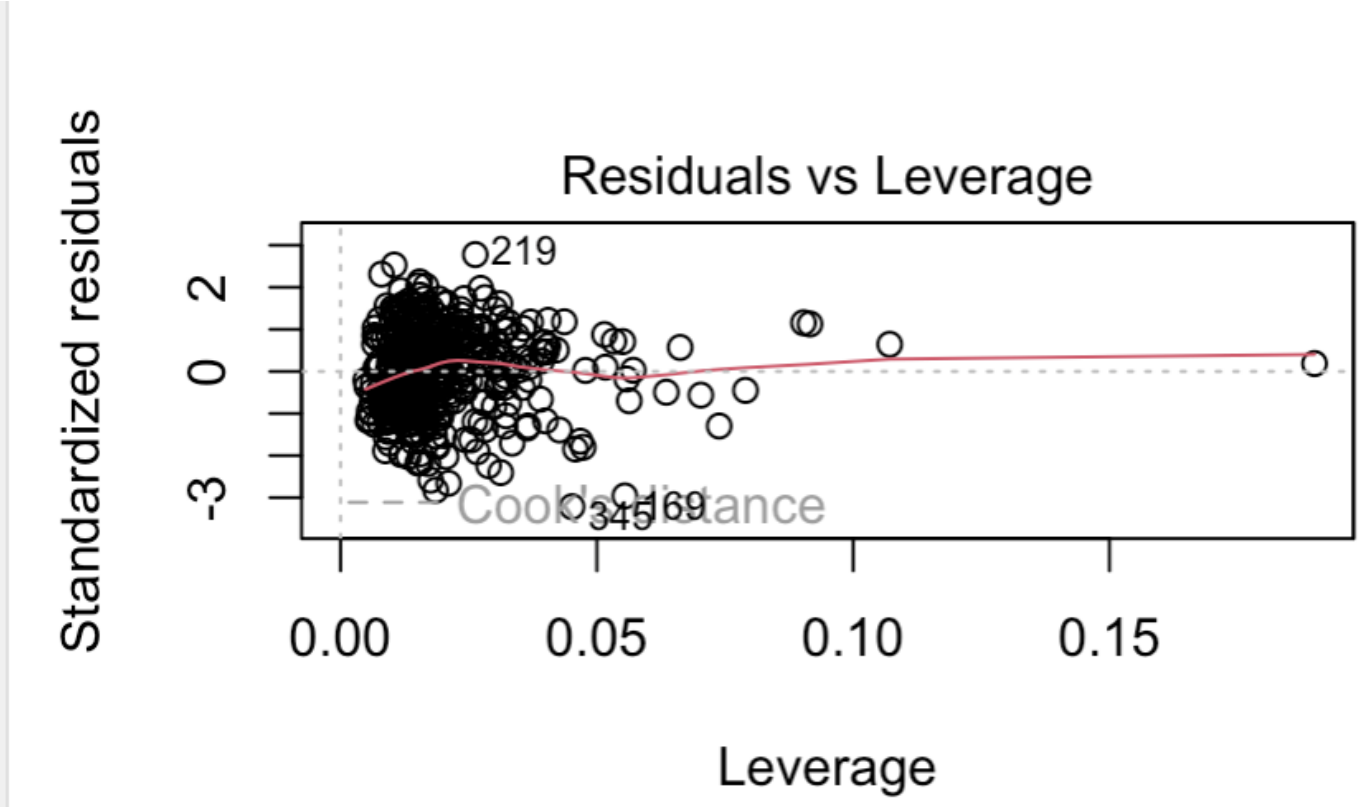


Figure 24 Cook's Distance

Figure 24 shows the cook's distance graph, for it is possible to conclude that there are no evident bad behaviors due to the outliers. However, to improve the verify this a deeper analysis has been executed:

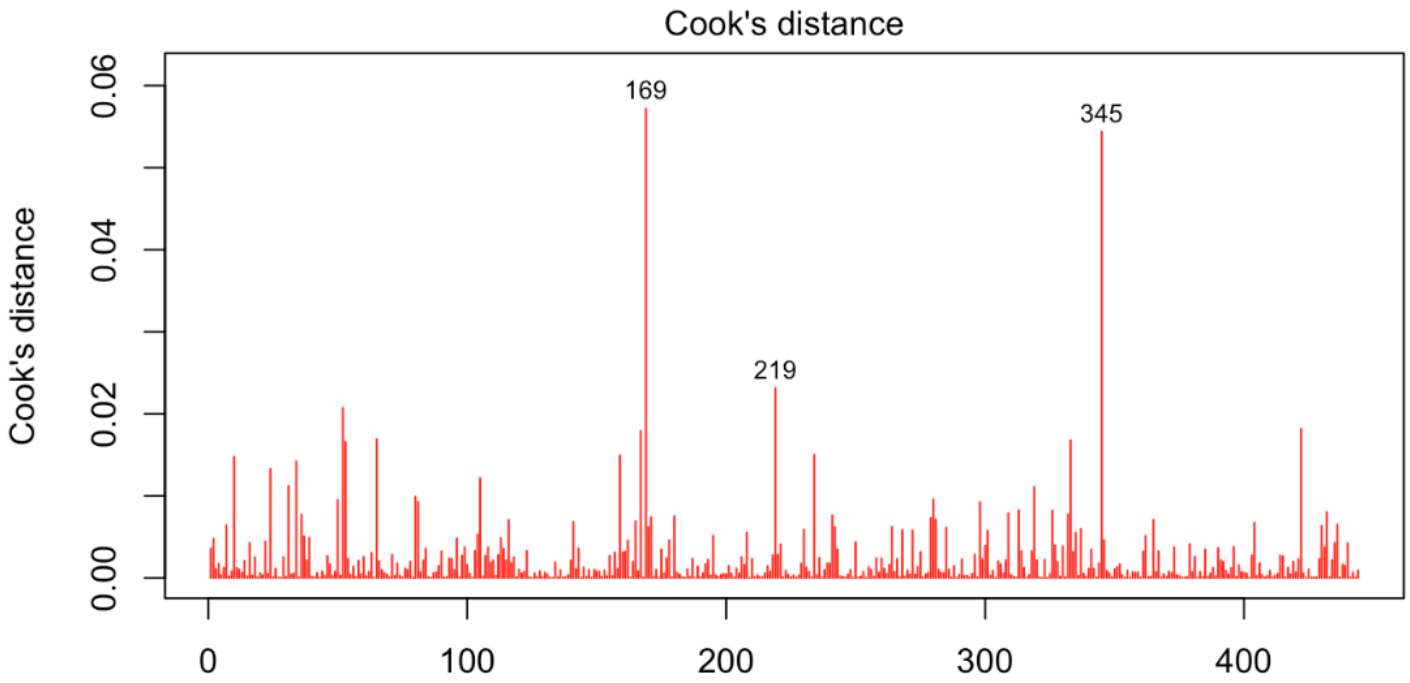


Figure 25 Cook's Distance Outliers

From figure 25 it is noticeable that there are no strange behaviors from outliers they are under the cook distance 0.06 and that is lower to our set 0.05 value. From this analysis we can conclude that there is no necessary to study the leverage points.

Analysis Part 2 - Coach Analysis:

1. Data Cleaning

The first step in our analysis was loading in the dataset and filtering for the 2022-23 and 2023-24 NBA seasons. Unfortunately, we were unable to combine these two seasons with the coach salary dataset due to coaches being in multiple years, so we decided to choose the season with the most accurate salaries. Since we had more coach salaries for the 2022-23 season than the 2023-24 season, we ended up using the 2022-23 year.

This is what the original dataset looks like (Figure 2.1):

Coach	Team	hs with Fra	ars as Coa	nt Season C	nt Season	nt Season	nchise Gan	nchise Wi	nchise Los	areer Gam	Career Win	Career Lose	eer Win R	ason Game	ason Wins	ason Loses	e Games (s	e Wins (Ps	e Loses (P	son Game	ason Wins	ason Loses
Lloyd Pierce	ATL	2	2	67	20	47	149	49	100	149	49	100	0.329									
Brad Stevens	BOS	7	7	72	48	24	564	318	246	564	318	246	0.564	17	10	7	73	37	36	73	37	36
Kenny Atkinson	BRK	4	4	62	28	34	308	118	190	308	118	190	0.383				5	1	4	5	1	4
Jacque Vaughn	BRK	1	4	10	7	3	10	7	3	226	65	161	0.288	4	0	4	4	0	4	4	0	4

Each column has been described in the first part of this report.

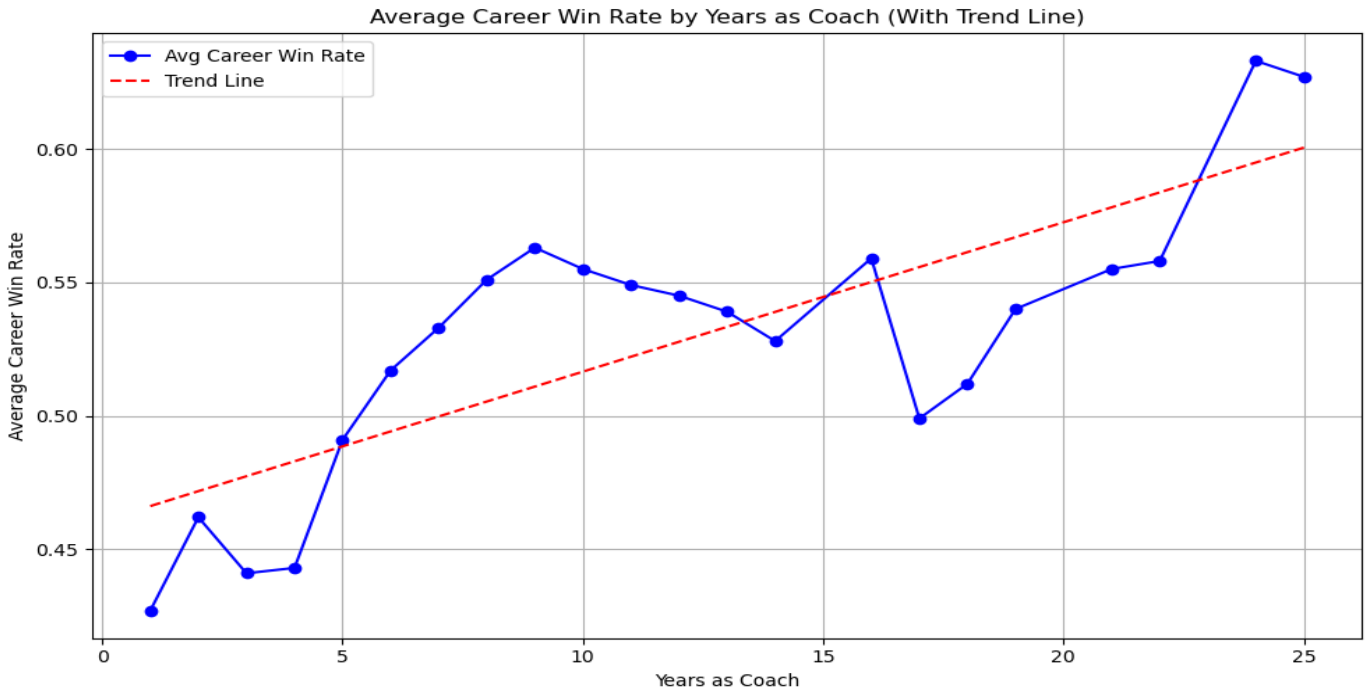
Before doing any analysis, these were the types of questions that we were interested in answering for coaches:

1. Do coaches with more relevant education get paid more?
2. Do coaches who previously played in the NBA get paid more?
3. Do coaches earn more as they gain more years of experience?

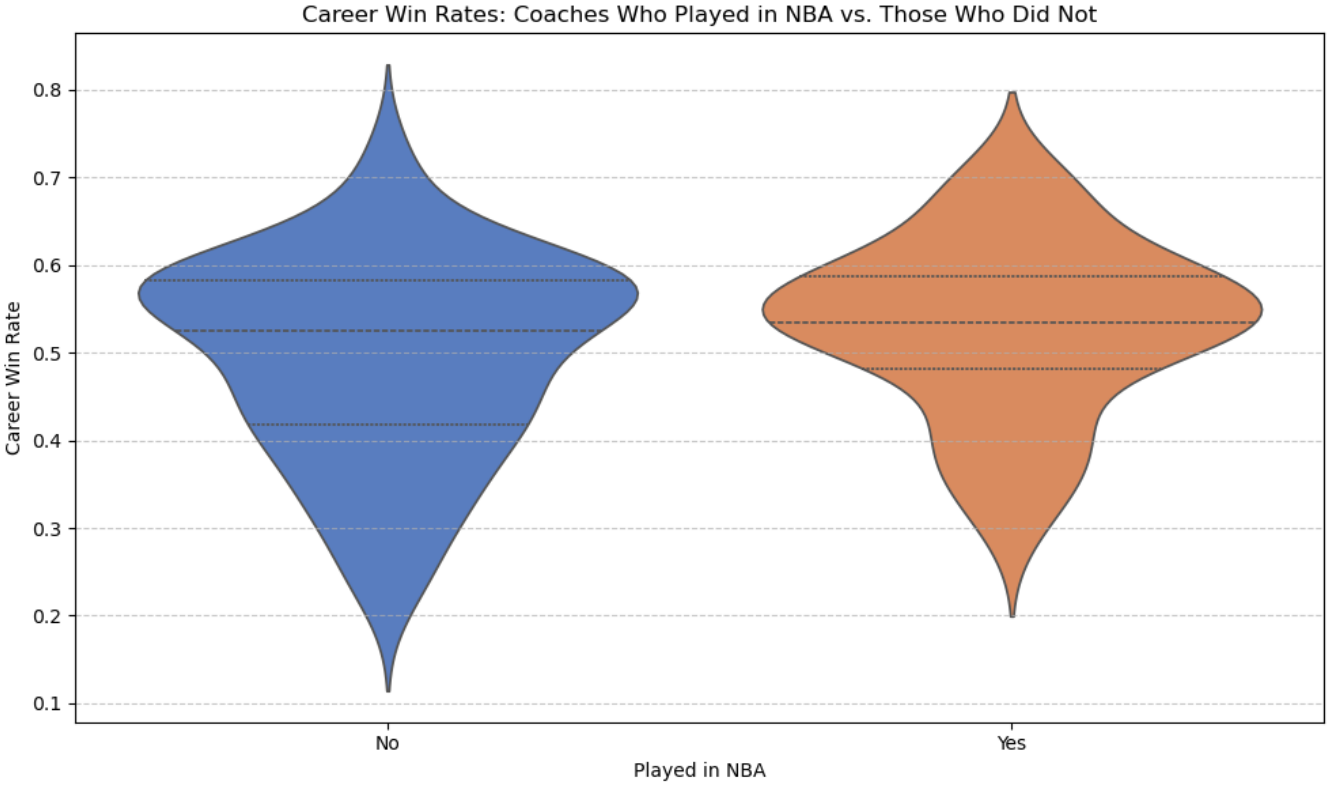
Before doing any analysis, we wanted to explore the data to see what types of relationships existed. Specifically, we wanted to see if there was any relationship between the demographic data and the on-court performance data like win

rates. We first graphed the experience of coaches and their average win rates, which ended up having a relatively strong looking relationship in the graph (Figure 2.2).

```
plt.show()
```



We were also interested in seeing if having previously played in the NBA had any advantages with being a better coach. Getting a better understanding of this would help us analyze if coaches with previous experience are actually better and may deserve to be paid more than coaches without experience. However, based on our graph we would expect that having played in the NBA does not seem to help coaches with success. We will keep this in mind during our analysis of coach salaries (Figure 2.3).



One of the most difficult aspects of analyzing coach salaries was that their salaries are confidential. Trying to track down each coach's salary was very difficult and most articles or wikipedia pages did not seem highly accurate. Keeping in mind that having poor data will end up creating a poor model, we did not spend too much time on our analysis. Once we found out just how difficult it was to find any salaries for coaches, we decided that we would continue doing analysis but focus most of our time on player analysis.

Our first additive model included all variables that would help us answer the questions highlighted above. We wanted to predict salaries since we thought it would give us an interesting perspective that may help with our analysis of NBA player salaries. It was also really interesting to see how much some teams are paid compared to the coach since the coach is typically assumed to have the most power on a team and can decide which players get to go on the court. We included years of experience since we assumed that like most career paths, as you gain more experience you have more negotiating power and can demand a higher salary. We also decided to include year of experience since we found it had a relationship with total wins meaning that coaches with more years of experience may have an advantage which therefore could be used to negotiate a higher salary.

One of the relationships we were most excited to learn more about was figuring out if coaches who played in the NBA when they were younger made more than non-NBA coaches. It can be very difficult to determine how good a coach is since not many coaches get to land assistant coach positions or head coach positions for top university teams. We initially thought that having experience playing in the NBA would be a huge advantage and they may earn more salary since they would already be so familiar with the game and would have proven their dedication by making it into the league. For example, one of the most decorated NBA coaches is Steve Kerr who coaches the Golden State Warriors. When he was a player in the NBA, he won five championships, which shows just how good of a player he was and how much knowledge he had in the game.

We also included the career win rate since we found that players who did better on court earned more money, so we assumed the same as a coach. No one should be paid more money if they consistently lose and are not able to make it into the championship round. We assumed that there would be a very strong relationship between actual performance as a coach and their salary. Finally, the last variable that we added was a fun relationship that we wanted to explore. We

determined that having previous education in areas like Physical Education and Business would be more advantageous than education in areas like American History or even Biology.

Our first additive model was this:
additive_model <- lm(Salary ~ `Years as Coach` + `Career Win Rate` + `Played in NBA` + Relevant_Education, data = basketball_clean)

Using a significance level of 0.05 and testing each variable with an individual t-test where the null hypothesis is that the specific variable is not significant (slope = 0) against the alternative hypothesis that the variable is significant (slope != 0). By going through each variable, we found that years as coach and having a relevant education were statistically significant with p-values of 3.0 e-4 and 0.02 (both less than 0.05).

However, we also wanted to test this model against an interaction model of the same variables:

interaction_model <- lm(Salary ~ (`Years as Coach` + `Career Win Rate` + `Played in NBA` + Relevant_Education)^2, data = basketball_clean)
Using a significance level of 0.05 and testing each variable with an individual t-test where the null hypothesis is that the specific variable is not significant (slope = 0) against the alternative hypothesis that the variable is significant (slope != 0). By going through each variable, we found that playing in the NBA, having relevant education, and all interaction variables were statistically significant since they had p-values of less than 0.05.

Now this is where we realize that we have some major problems with our dataset. Of all the variables in our dataset, years as a coach and career win rate should be the two most statistically significant variables. We had concerns going into the analysis of coach salaries since the data was so difficult to collect and we used so few coaches to build the model.

We then cut out all statistically insignificant variables from our models. Due to the hierarchical rule, since all our interaction variables were significant, we could not get rid of years as coach and career win rate. After running our new models again, our refined additive model had all statistically significant variables (p-values less than 0.05) however our interactive model needed to have played in NBA removed and both interactive variables including played in NBA removed as well (Figure 2.4).

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        5276440    2273583   2.321 0.027049 *
`Years as Coach`                  -1104468     396884  -2.783 0.009094 **
`Career Win Rate`                  3334009    4012956   0.831 0.412434
`Played in NBA`Yes                   26971     877766   0.031 0.975684
Relevant_Education1                20555744    4843307   4.244 0.000185 ***
`Years as Coach`:`Career Win Rate`  1973503     626858   3.148 0.003619 **
`Years as Coach`:`Played in NBA`Yes   -15480      73446  -0.211 0.834446
`Career Win Rate`:Relevant_Education1 -36837041    8182311  -4.502  8.9e-05 ***
`Played in NBA`Yes:Relevant_Education1   728356    1328447   0.548 0.587429
---
```

This time, our interactive model had all significant variables (Figure 2.5).

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       5310378    1996443   2.660 0.011837 *
`Years as Coach`                 -1107581     294084  -3.766 0.000629 ***
`Career Win Rate`                 3184571    3766782   0.845 0.403778
Relevant_Education1               21200579    4315192   4.913 2.23e-05 ***
`Years as Coach`:`Career Win Rate`  1976266     490681   4.028 0.000299 ***
`Career Win Rate`:Relevant_Education1 -37680653    7338597  -5.135 1.15e-05 ***
---
```

We then conducted an anova test to determine if the model with interactive variables was a better fit. We tested this using a null hypothesis that the additive model would be a better fit against the null hypothesis that the interaction model should be used. With a p-value of 1.78 e-5, we reject the null hypothesis and conclude that the model with interactive variables is better.

In order to confirm our conclusion, we also compared the adjusted r squared and RSE values of the models (Figure 2.6).

Model <chr>	Adjusted_R2 <dbl>	RSE <dbl>
Refined Additive	0.3491861	1692304
Refined Interaction	0.6246279	1285231
Final Interaction	0.6542087	1233551

As we can see, the final interaction model has the best adjusted R squared value and the lowest RSE which proves that it is the best model we have tested.

The final model for predicting coach salaries is
*Predicted Salary = 5310378 - 1107581 * `Years as Coach` + 3184571 * `Career Win Rate` + 21200579 * Relevant_Education1 + 1976266 * `Years as Coach`:`Career Win Rate` - 37680653 * `Career Win Rate`:Relevant_Education1 where the coach has relevant education*

*Predicted Salary = 5310378 - 1107581 * `Years as Coach` + 3184571 * `Career Win Rate` + 1976266 * `Years as Coach`:`Career Win Rate` where the coach does not have relevant education*

Since this model only accurately predicts 65% of the variability in coach salaries, it clearly indicates that this is not a strong model and we are missing many variables that are relevant.

4 CONCLUSION AND DISCUSSION

4.1 Approach

Our approach looks at the different factors that influence NBA player salaries using multiple regression models, including interactions between variables. We focused on things like age, performance (points, assists), and draft position to see how they work together to affect salaries. This method goes beyond simple relationships and shows how these factors can change in importance when combined, like how age impacts the value of scoring or playmaking.

Using interaction terms like age × points and age × assists was a good decision because it shows that the relationship between a player's characteristics and their salary isn't always simple and can change depending on things like age or career stage. This approach gave us a better fit and more accurate results (adjusted $R^2 = 0.7698$) compared to simpler models, making it a strong method for predicting salaries.

To improve the model, we could:

1. Explore time-series models, as salaries change over seasons and can be influenced by factors like contract length or market trends. If our data covered multiple seasons, we could help capture season-to-season patterns, adding a time-based perspective to the analysis.
2. Use a hierarchical model to improve the analysis by accounting for the fact that players from the same team or season might share certain characteristics.
3. Use clustering techniques to group players based on similar characteristics like age, height position, and performance. Then, we can predict salaries within these clusters, which might help capture more modified salary

trends.

4. Create new features by combining important metrics like points per game, assists per game, and rebounds per game, and categorize them under variables such as offensive abilities

4.2 Future Work

To improve this analysis, the following follow-up work could be done:

1. Use of more accurate salary data for both player and coach datasets.
2. Incorporate Time-Series Models: Time-series models like ARIMA or LSTM could help us understand how salaries change over time. These models would allow us to capture patterns from season to season, considering how market trends impact salaries.
3. Expand the Dataset: We could improve the analysis by adding more data, such as player injuries, contract details, or team performance. This would help us understand how these factors influence salaries beyond just performance stats.
4. Creating new features, like adjusting performance stats for a player's age or career stage, could help make the model more accurate. This would give us a better understanding of how a player's experience impacts their salary.
5. Cross-Validation: Using cross-validation will help test the model on different sets of data, making sure it performs well not just on the training data but also on new data. This helps prevent overfitting and ensures the model is more reliable.

By working on these areas, the analysis can become more accurate and get better insights into the factors that affect NBA player salaries.

Coach Analysis:

With only 65% of variation in coach salary being explained by our final model I would say that we had varying success with this analysis. We ran into major problems with data integrity issues due to the lack of accurate information and using so few coach data points. Looking at our three questions that we wanted to analyize further: Do coaches with more relevant education get paid more? Do coaches who previously played in the NBA get paid more? Do coaches earn more as they gain more years of experience?

Using our final model, we can conclude that based on our analysis, it does not matter if a coach played in the NBA or not when they were younger. Both models found that NBA experience was statistically insignificant which makes sense since coaching uses completely different skills than playing in the NBA. This might be explained since performance of the coach may be much more important than any other demographic variable. When we look at the impact of having relevant education, it all depends on the coaches career win rate. What does not make too much sense is that when we hold all variables constant and compare having relevant education vs nonrelevant education with different win rates. Coaches with relevant education and lower win rates on average will earn more than coaches with relevant education and a high win rate using our model. You would expect that coaches of the same education level would be paid more if they performed better. This indicates that our model may not be very accurate. Years of experience is positively correlated with earning more which makes logical sense.

The final model for predicting coach salaries is
*Predicted Salary = 5310378 - 1107581 * `Years as Coach` + 3184571 * `Career Win Rate` + 21200579 * Relevant_Education1 + 1976266 * `Years as Coach`:`Career Win Rate` - 37680653 * `Career Win Rate`:Relevant_Education1 where the coach has relevant education*

*Predicted Salary = 5310378 - 1107581 * `Years as Coach` + 3184571 * `Career Win Rate` + 1976266 * `Years as Coach`:`Career Win Rate` where the coach does not have relevant education*

In the future, it would be interesting to revisit this analysis and use more accurate data. We would expect to find better results if we used more years of data however the most improvement would come from finding more accurate salary information for coaches.

One question that would be very interesting to explore in the future would be if coaches on teams with higher budgets get paid more than coaches on teams with lower budgets. Although it would be extremely difficult, it would also be really interesting to see how the impact of social media influences salaries in both player and coach datasets. We could add in a new column that has the total followers of each coach and player on instagram and use that as a predictor of salary.

In conclusion, we found that it does not matter if a coach played in the NBA or not, coaches with more experience will earn more, and there seems to be a relationship between relevant education and salary that we have not been able to fully predict accurately. Although our final model is not very accurate, it gives us a good understanding of where to continue our analysis in the future and relies heavily on accurate data which is not currently available.

Players Salary Analysis:

The model that has been selected is show in the figure below.

$$\widehat{Salary} = \hat{\beta}_0 + \hat{\beta}_1 X_{age} + \hat{\beta}_2 X_{playerHeight} + \hat{\beta}_3 X_{gp} + \hat{\beta}_4 X_{pts} + \hat{\beta}_5 X_{ast} + \hat{\beta}_6 X_{I(age^2)} + \hat{\beta}_7 X_{draftNumber} + \hat{\beta}_8 X_{draftNumber} * X_{pts}$$

Figure 26 Final Proposed Model

From figure 26, it is possible to see the model proposal, figure 27 shows the summary of the propouse model:

```
Call:
lm(formula = log(salary) ~ age + player_height + gp + pts + I(age^2) +
    ast + draft_number + pts * draft_number, data = player_salaries_2020_21)

Residuals:
     Min       1Q   Median       3Q      Max
-1.93447 -0.43556  0.06826  0.43396  1.68722

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -7.305019   1.296418  -5.635 3.15e-08 ***
age               0.293966   0.077452   3.795 0.000168 ***
player_height     0.018163   0.003804   4.775 2.45e-06 ***
gp                0.001477   0.001829   0.808 0.419613
pts               0.040078   0.010256   3.908 0.000108 ***
I(age^2)         -0.003757   0.001402  -2.680 0.007633 **
ast               0.115392   0.025017   4.613 5.24e-06 ***
draft_number     -0.339102   0.035768  -9.481  < 2e-16 ***
pts:draft_number  0.012472   0.003682   3.387 0.000770 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6155 on 435 degrees of freedom
Multiple R-squared:  0.724,     Adjusted R-squared:  0.7189
F-statistic: 142.6 on 8 and 435 DF,  p-value: < 2.2e-16
```

Figure 27 Purpose Model Summary

From the summary, it is possible to obtain the coefficients for the model:

- Intercept: -7.305: This is the expected salary when all predictor variables are zero
- age: 0.294: For each one-year increase in age, the salary increases by approximately 0.294 units, holding other factors constant.
- player_height: 0.018: For each centimeter increase in height, the salary increases by 0.018 units.
- gp (games played): 0.001: For each additional game played, the salary increases by 0.001 units.
- pts (points per game): 0.040: For each additional point scored per game, the salary increases by 0.040 units.
- I(age^2): -0.004: The squared age term is included to capture any non-linear relationship between age and salary. Here, a negative coefficient suggests diminishing returns of age on salary at higher ages.

- ast (assists per game): 0.115: For each additional assist per game, the salary increases by 0.115 units.

- draft_number: -0.339: For each increase in draft number (i.e., a worse draft position), the salary decreases by 0.339 units.

- pts:draft_number (interaction term): 0.012: This indicates the combined effect of points per game and draft number on salary. For each unit increase in the product of points per game and draft number, the salary increases by 0.012 units.

In conclusion the proposed model has a R adjusted of 71.89% this means that the model studies 71.89 percent for the behavior of the salary in this dataset, having a categorical grouping for the drafts and applying log transformation to have a normal data. The model meets all the assumption but the heteroscedasticity, this requires a deeper analysis.

5 REFERENCES

1. Mark W. Sanchez, “Steve Kerr admits ‘I’m hungover right now’ day after Warriors’ championship”, 06/2022, Publisher, NBA. https://nypost.com/2022/06/18/steve-kerr-admits-im-hungover-right-now-after-warriors-title/

2. Michael H. “NBA Team stats”, Publisher, kaggle https://www.kaggle.com/datasets/mharvnek/nba-team-stats-00-to-18/code.
3. Justinas Cirtautas. “NBA Players” Publisher, kaggle . https://www.kaggle.com/datasets/justinas/nba-players-data/data

4. Team dunkests, “How Much Is a NBA Championship Ring Worth” Publisher.Dunkest. https://www.dunkest.com/en/nba/news/127624/how-much-nba-championship-ring-worth,

5. NBA hoopsnype "Salaries” Publisher. https://www.kaggle.com/datasets/justinas/nba-players-data/data
6. Chris Schwegler, “Shots from the Court: Pistons vs Cavaliers”, 11/2024, Publisher, NBA .https://www.nba.com/pistons/photos/shots-from-the-court-pistons-vs-cavaliers

End of Project Report