STA302 Fall 2023 Methods of Data Analysis 1
Final Project Proposal (Part 1)

Names of Group Members	Contribution to Proposal
Alex Lugard	A3, A4, C2, C3, C4 + Overall Editing
James Li	A1, A2, A3, A4 + Overall Editing
Tristan Gauntley	A3, A4, B1, C1, C2, C4, C5 + Code & Overall Editing
Ziana Suleman	A1, A2, B1, B2, B3, + Code & Overall Editing

A. Research Question and Supporting Literature

1. *What is the research question you will be studying in this project? Be sure to explicitly refer to the variables under study and avoid using vague language to describe your study question.*

What effect do Number of Games Started, Turnovers per Game, Field Goals Attempted per Game, Age, and a player's Primary Position have on an NBA player's average Points per Game?

2. *Provide an explanation for why a linear regression model would allow you to answer your research question. What aspect of your fitted model would give you the answer.*

A linear regression model would allow us to answer the effect of certain statistics on an NBA player's average Points per Game because it will allow us to visualize the strength of the relationship between each predictor variable and our response variable. Specifically, we will be able to use data from different NBA players to visualize how the number of points an NBA player scores in a game will change for a one unit increase of each predictor variable: Number of Games Started, Turnovers, Field Goals Attempted, Age, and Primary Position. Since primary position will be a categorical variable, it will determine how the points scored by an NBA player is affected by their specific position.

3. *Provide proper citations for 3 peer-reviewed academic research articles related to your specific research question or your topic of interest. For each, describe how the results of the article relate to your research question. Further, rank each article on a scale of 1 to 3 (1=not useful, 2=slightly useful, 3=very useful) based on how useful the article is in providing insight into the population relationship you wish to estimate. Justify this ranking.*

Citation	Description, Ranking and Justification
Anton Kalén, Alexandra Pérez-Ferreirós, Pablo B. Costa & Ezequiel Rey (2021). "Effects of age on physical and technical performance in National Basketball Association (NBA) players"DOI: 10.1080/15438627.2020.1809411	**(Rank = 2)** This paper is a study on the effect of age on a player's physical and technical performance (one aspect being points scored). The ranking is "slightly useful" as the article analyzes the statistical correlation between age and points scored, which is one of the questions we seek to answer in this project, but does not use linear regression.
Papadaki, Ioanna & Tsagris, Michail. (2022). "Are NBA Players' Salaries in Accordance with Their Performance on Court?". 10.1007/978-3-030-85254-2_25. https://www.researchgate.net/publication/357883185_Are_NBA_Players%27_Sala	**(Rank = 2)** This paper is an analysis of the correlation between salary and their performance statistics as predictor variables (APG, RPG, PPG). However, it has been determined that these statistics are non-linear, requiring non-linear regression to model the relationships. As such, this article is ranked "slightly useful" since the objective of the study is similar to the research question posed but rather non-linear regression is used to achieve its conclusion.

ries_in_Accordance_with_Their_Performance_on_Court	
Page, Garritt & Barney, Bradley & McGuire, Aaron. (2013). "Effect of position, usage rate, and per game minutes played on NBA player production curves". Journal of Quantitative Analysis in Sports. 1-9. 10.1515/jqas-2012-0023.	**(Rank = 2)** This paper utilizes John Hollinger's usage rate statistic and average minutes played to produce individual production curves using a Gaussian Process regression. In addition, this paper aims to model the production of the player throughout his whole career rather than his "prime". The ranking given is "slightly useful" as we would like to determine a correlation between MPG and production using a Gaussian Process regression instead of a linear regression model, and it does not predict year-by-year but seeks to model a curve throughout the player's career.

4. *Provide the database/library where you located the above academic papers. List the search terms used to find these papers, in addition to the number of results for each search term.*

Database/library searched	Search terms used	Number of results for each
Google Scholar — PubMed, Association for Computing Machinery (ACM)	NBA, NBA PPG, NBA linear regression	On Scholar — 335000, 5600, 15900 searches, respectively
Research Gate	NBA Players Performance	On Research Gate — 100 searches
Research Gate	Effect of position on NBA player	On Research Gate — 100 searches

B. Data Description, Justifications and Summary

1. *Provide the website from which your chosen data was obtained/downloaded.*

Website:**	https://www.kaggle.com/datasets/davra98/nba-players-20162019

*** If your data was obtained from a data repository (e.g. Kaggle, UCI Repository, etc.), please state how your research question differs from the original purpose of this data.*

The data was obtained from Kaggle. The original purpose of the data, explained by the author, was to create a model that predicted the score of each NBA player in the all-star game. Our research question differs from this because we are observing how different factors (predictor variables) will impact the PPG of an NBA player throughout the season.

2. *List the variables you have selected to be part of your preliminary model (minimum of 5 with at least one a categorical variable). Please give an understandable name to each variable rather than writing the name that appears in R. For each variable, justify why you have chosen to use this variable over others in the dataset, and what the role of each variable will be (e.g., predictor of interest, predictor informed by literature, confounder, etc.).*

Variable Name	Justification for Use	Role in Model
Points per Game (PPG)	PPG was chosen as the response variable because it will be directly affected by each of the predictor variables. It also helps to evaluate the scoring performance of a player. If a player is scoring more points, that means that they are a valuable asset to the team and are making a positive contribution towards winning. This was chosen over other variables in the dataset as other variables more often affect points per game, rather than are influenced by points per game. Furthermore, points per game is oftentimes regarded as a key statistic in the general performance of a player, and is therefore interesting to analyze as a response variable.	Response Variable
Games Started	A team's starting lineup generally consists of the players who perform the best on the team. This means that the more games a player starts, the better player they are, and hence the more points they should score, regardless if they are a lower scoring player, they would still be expected to score more than a bench player. This was chosen over variables such as minutes played in the dataset because sometimes non-starters are given time on the court to rest starting players or if a team has a large lead towards the end of the game they will rest their best players because they know they cannot lose. Number of games started is a good reflection of the best players on the team who should score more points the more games that they start.	Predictor Variable (Confounder)
Turnovers	If a player has a higher number of turnovers in a game, they are taking away a chance for their team (and themselves) to score which will reduce their points per game. This was chosen over other variables in the dataset, especially steals, since an increase in turnovers implies that the player has the ball in their hands more often, which also means that same player will tend to score the ball more. Although steals may lead to an increase in fastbreaks, steals often lead to diverse situations where the player initiates the fastbreak and may lead to assists instead. Thus, turnovers give us a better predictor-response relation in this case rather than steals.	Predictor Variable (Predictor of Interest)
Field Goal Attempts	Players that have a higher number of field goal attempts are going to have a higher chance of scoring more than players with a lower number of field goal attempts. This was chosen over field goal percentage because shooting a higher percentage of field goals will have less of an impact on PPG compared to attempting more field goals. For example, if a player has a FG%=50% scoring 2/4 shots and another player has a FG% of 40% but made 4/10 shots, they will have scored more points for their team.	Predictor Variable (Confounder)
Age	According to our literature review, "Effects of age on physical and technical performance in National Basketball Association (NBA) players", as a player gets older, their skills will develop leading to an increase in PPG. Younger players have just been drafted and are still developing their skills therefore they will have a lower number of PPG. At a certain age, players reach their production value and after this their skillset decreases as they get older due to the decrease in physical performance. This was chosen over other variables that are not performance based statistics of an NBA player such as salary because salary does not change points per game but PPG could change salary which shows that it is not a predictor variable.	Predictor Variable (Predictor Informed by Literature)
Primary Position	Five players start in every NBA game and each player holds a different position which will be observed as a categorical predictor variable where the numbers 1-5 will represent a different position. This was chosen because certain positions dictate different jobs of a player on a team. For example, the main goal of a shooting guard is to score whereas the main goal of a center is to get rebounds, block shots, etc. As position changes to players whose main job it is to score, the PPG response should reflect this. We chose the primary position categorical variable over the team categorical variable because position will affect points per game more than the name of the team since the name of a team does not affect the PPG of a player.	Categorical Predictor Variable (Predictor of Interest)

3. *Produce a table of numerical summaries of the variables listed above. Summaries should be appropriate to the type of variable, and interesting/important characteristics about variables should be mentioned in an informative caption. Include your summary table below.*

Variable and Informative Caption	**Numerical Summary**						
	Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**	**Standard Deviation**
Points per Game: The summaries show that there is a large range of points and the maximum value is much larger than the value in the 3rd quartile showing that fewer players have very high PPG.	0.000	4.300	7.300	8.794	12.00	36.10	6.082
Games Started: The value of games started shows that the average player only starts in approximately 30% of games (max is 82) potentially affecting the quality of data since we are considering a wide range of players.	0.000	1.000	13.00	25.91	52.00	82.00	28.66
Field Goal Attempts: Most players (under the 3rd quartile) attempt less than 10 field goals in a game which will overall lead to lower PPG from players.	0.000	3.700	6.150	7.173	9.900	24.50	4.592
Age: The NBA overall consists of players who are in their 20s and very few players exceed the age of 30. The wide range of players in the data set provides an accurate estimation of the data relevant to age.	19.00	23.00	25.00	26.14	29.00	42.00	4.283
Turnovers: Most players have a very low number of turnovers, which could be a result of quality of the player, or lack of playtime.	0.000	0.600	0.900	1.131	1.500	5.700	0.8032

Categorical Variable and informative caption	**Numerical Summary**				
Position: Ideally the data set would include equal amounts of each position, but from the table, it is evident that more shooting guards are included compared to small forwards. This may have an effect on the accuracy of the data regarding the trends in position and points.	C: 277	PF: 295	PG: 282	SF: 230	SG: 324

C. Preliminary Model Results

1. *Fit your preliminary multiple linear model and present the estimated relationship. Present this information carefully so that it is easily readable and understandable.*

```
Call:
lm(formula = new_data$PTS ~ new_data$Age + new_data$TOV + new_data$GS +
    new_data$FGA + new_data$Pos1)

Coefficients:
    (Intercept)     new_data$Age     new_data$TOV      new_data$GS
      -0.178819         0.010921         0.366318         0.004283
   new_data$FGA  new_data$Pos1PF  new_data$Pos1PG  new_data$Pos1SF
       1.241000        -0.645645        -1.142013        -0.773393
new_data$Pos1SG
      -1.082634
```

2. *Justify your choice of how you included the categorical variable in your preliminary model. How does this choice contribute to answering your research question?*

Our choice for the categorical variable is Primary Position. Depending on a player's position, they are more skilled/encouraged to shoot or make a play on the basket. For example, a point guard is trained to have a good shot and therefore will more likely score points as opposed to a power forward, whose skills are defending and rebounding (both non-scoring skills). Therefore, it is of interest to plot and analyze the effect of Primary Position on Points Per Game.

3. *Do your estimated coefficients align/agree with the results of your three peer-reviewed articles? Explain in what way they differ/agree and provide a reason why this might be the case.*

Our coefficients agree with some of the findings of the three peer-reviewed articles, but disagree with others. Specifically, our results show age having a small positive correlation with points scored, which agrees with the article focusing on this variable. According to the article, "technical performance generally remained stable or increased [with age]". (Anton Kalén) While it is not explicitly stated that turnovers cause a decrease in points, we can interpolate from the second peer-reviewed article (Papadaki) that since turnovers decrease the performance of a player, this implies less points scored. However, our data disagrees with this conclusion. This is most likely due to the fact that our data doesn't take into account who is in control of the ball and takes most of the shots. For example, because point guards carry the ball most of the time, they are bound to also cause the most turnovers, while simultaneously scoring many points, hence skewing the data. On the contrary, our strong correlation between field goal attempts and points scored strongly agrees with the conclusion of this article. In order to tell whether our results for games started agrees with the article, we need to infer certain results from others in the article. The paper clearly states that minutes played is correlated with points scored. From this we can confidently infer that games started would similarly affect points scored, since games started and minutes played are heavily correlated. Hence, our data agrees with this conclusion. Finally, the correlation between position and points scored neither disagrees or agrees with the final article chosen (Page). In fact, the article emphasizes the fact that regardless of position, it is the players with high minutes that tend to have greater fluctuation in game score (points scored). Our results may therefore be inconclusive, as the data set includes many players that play little to no minutes, reinforced by the very low mean number of games started (25.91). The data set also shows an average points per game of 8.794, and paired alongside the low number of games started, indicates that the data contains a higher quantity of bench players as opposed to superstars, where the points per game of the player may follow a different trend than what we sought out.

4. *Perform a complete assessment of the assumptions of your preliminary model. Do you observe violations of assumptions or conditions? Describe how you came to this conclusion, making explicit reference to any plots or other information that is relevant.*

1) Uncorrelated Errors Assumption
This assumption is *not violated*. We arrive at this conclusion since there isn't a clear large cluster of data in the "Residual vs. Fitted" plot (Figure 1). The fanning pattern is accounted for in the violation of the constant variance assumption, and there are no other clear systematic patterns on any of the residual vs. predictor plots (Figure 2). Therefore, this assumption is not violated.
2) Linearity/Mean Zero Error Assumption
This assumption is *not violated*. We arrive at this conclusion by examining both the "Residual vs. Fitted" plot (Figure 1) and the residual vs. predictor plots (Figure 2)
3) Constant Variance Assumption
This assumption is *violated*. We arrive at this conclusion by examining the "Residual vs. Fitted" plot (Figure 1), in which we see a fanning pattern of the residuals as the fitted values increase. The spread clearly widens, thus indicating that the constant variance assumption is violated. Furthermore, the "Residual vs Turnovers" graph, and "Residual vs Field Goals Attempted" plots (Figure 2) both show a clear fanning pattern and hence this assumption is indeed violated.
4) Normality Assumption
The normality assumption is *violated*. This can be seen from Figure 4, where theoretical quantiles are plotted versus sample quantiles. If this assumption were not violated, the graph would look almost completely linear, with a clear one to one ratio. However, our graph shows data trailing off at smaller and larger quantiles.
Condition 1: Conditional Mean Response
This condition *holds.* We arrive at this conclusion by examining the "Response vs. Fitted" plot (Figure 5), in which we see a clear linear trend of the data points around the plotted line. This indicates that the mean responses are indeed a single function of a linear combination involving the predictors.
Condition 2: Conditional Mean Predictors
This condition *holds.* We arrive at this conclusion by examining the pairwise scatterplots (Figure 6), in which we see no curves or any non-linear patterns in any of the plots. This indicates that the mean of each predictor is indeed related to each other predictor either linearly or with no relationship whatsoever.

After a complete assessment of the assumptions, it is clear that the preliminary model does not align with some of the critical assumptions necessary for linear regression.

5. *Include all relevant plots created for assessing model assumptions below, with appropriate axis labels and captions.*

Figure 1: Residual vs. Fitted

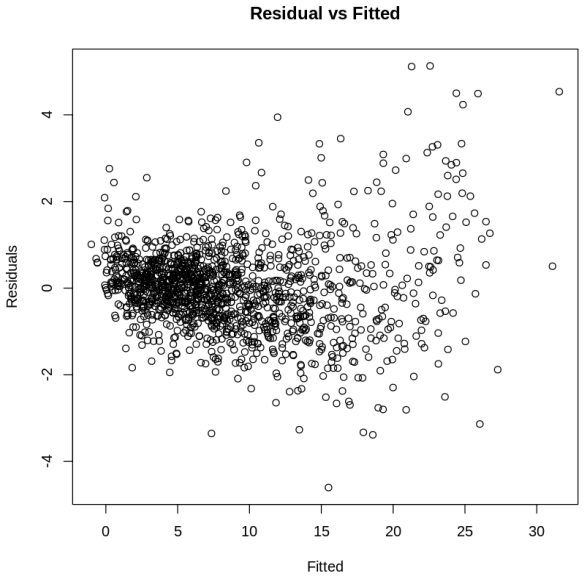

Figure 2: Residual vs. Predictors

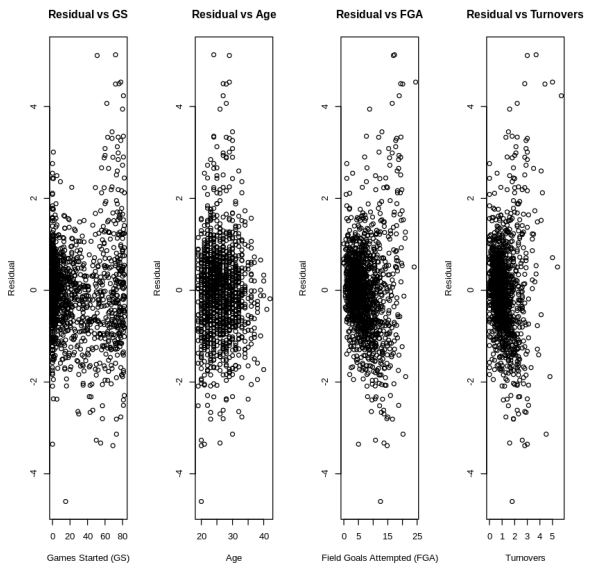

Figure 3: Residual vs Primary Position

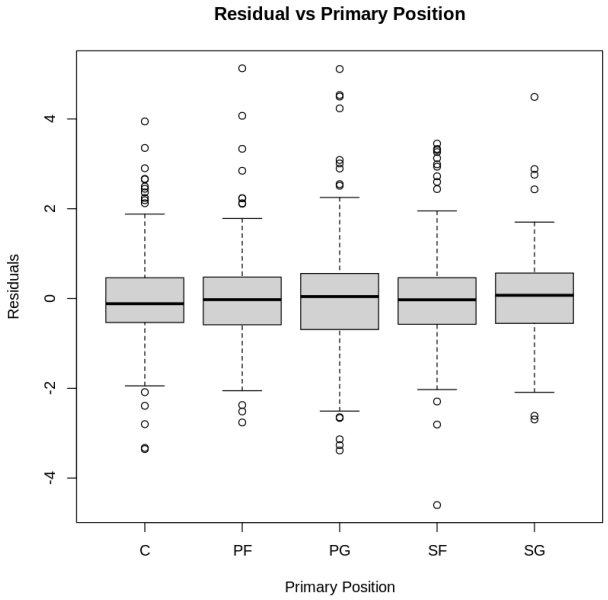

Figure 4: Normal Q-Q plot

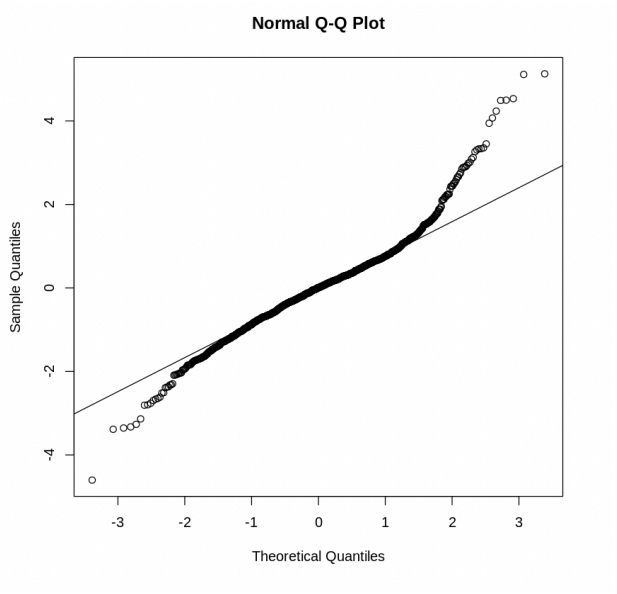

Figure 5: Response vs Fitted

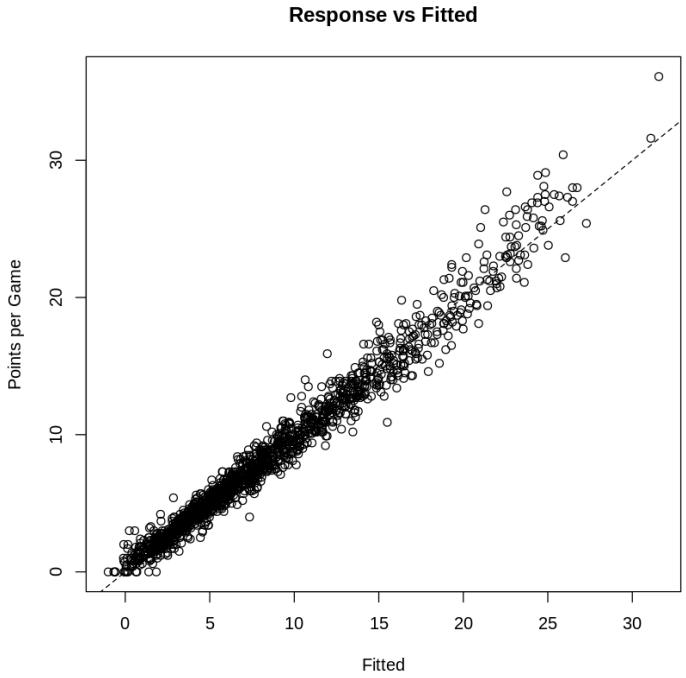

Figure 6: Pairwise Scatter Plot