Data Source (URL web address with hyperlink):

https://www.basketball-reference.com/leagues/NBA_2021_per_game.html

Context of Data and Variables:

The data is about 2020-21 NBA Player Stats， which contains name of player , position(Pos), age(Age), total games they played(G)， minutes they played per game(MP),field goal attempts per game(FGA),free throw attempts per game(FTA),assists per game(AST) and points per game(PTS).I'll regard Pos、Age、G、MP、FGA、FTA、AST and PTS as variables.

Output from str() function applied to the data object (apply a monospaced font like "Courier New" to the output):

```
> str(playerstats)
tibble [705 x 10] (S3: tbl_df/tbl/data.frame)
 $ rank  : int [1:705] 1 2 3 4 5 6 7 8 9 10 ...
 $ Player: chr [1:705] "Stephen Curry" "Bradley Beal" "Damian Lil
lard" "Joel Embiid" ...
 $ Pos   : chr [1:705] "PG" "SG" "PG" "C" ...
 $ Age   : num [1:705] 32 27 30 26 26 21 25 20 32 28 ...
 $ G     : num [1:705] 63 60 67 51 61 66 58 61 35 54 ...
 $ MP    : num [1:705] 34.2 35.8 35.8 31.1 33 34.3 35.1 33.2 33.1
 34.9 ...
 $ FGA   : num [1:705] 21.7 23 19.9 17.6 18 20.5 19.4 17 17.2 20.
1 ...
 $ FTA   : num [1:705] 6.3 7.7 7.2 10.7 9.5 7.1 5.1 8.7 6.8 4 ...
 $ AST   : num [1:705] 5.8 4.4 7.5 2.8 5.9 8.6 4.9 3.7 5.6 6 ...
 $ PTS   : num [1:705] 32 31.3 28.8 28.5 28.1 27.7 27.4 27 26.9 2
6.9 ...
```

Research Questions to be explore:

1.Is there any relationship between the PTS and the other seven variables?

2.Is there some interaction between variables?And are some variables proper to include to analyse?

Statistical Analysis Plan

Population

2020-21 Basketball Players in NBA.

Primary Objective:

To investigate the relationship between players’ scores per game and the other related stats, such as their position、age、total games played and so on.

Secondary Objectives:

Include interaction between variables and the other seasons stats to compare and analyse.

Data Collection methods:

Scraping all related data from website and choose some variables to analyse.

Variables under consideration:

PTS(continuous variable)—points the player scored per game—Response variable

Pos(categorical variable)—player’s position in the game which includes C、PF、SF、SG、PG—Explanatory variable

Age(continuous variable)—player's age on February 1 of the season 2020-21—Explanatory variable

G(continuous variable)—total games they played in the season—Explanatory variable

MP(continuous variable)—minutes they played per game—Explanatory variable

FGA(continuous variable)—field goal attempts per game .In other words, the number of times the play shot per game—Explanatory variable

FTA(continuous variable)—free throw attempts(penalty shot) per game—Explanatory variable

AST (continuous variable)—assists per game Explanatory variable

Missing data procedures:

There is no missing data. All data of players participate in season 2020-2021 are collected.

Numerical and graphical summaries to be presented:

Create boxplots of different variables according to different position.

Mean, quantile and standard deviation for seven continuous variables.

Scatterplot of every two variables find relationship or correlation between them to identify high correlations, and possible multicollinearity between explanatory variables, and identify explanatory variables we think could play an important role in our regression model.

Models to be fitted:

In terms of correlation or use forward and back forward selection methods to choose appropriate explanatory or interactions to include in our linear regression model.

General linear model:

PTS~ Pos + Age +G + MP + FGA + FTA +AST

PTS~ Pos+ Age+ G + MP + FGA + FTA

+AST+G:FGA+G:FTA+G:FTA+G:MP+MP:FGA+MP:FTA

Fit full model and then test to see if possible to simplify by dropping interaction term, significance set at 5%. Check residuals after final model for constant variance, normality.