Season's Informaiton and Post Injury Play Time For NFL Players

Group D: Arthur, Donna, Kaitlyn, Michael

Friday, April 6, 2024

Project Question/Aim

Concussions have been a trending topic in football for the last few years as current and former players suffer the consequences of such a physical sport. We want to determine if there is a significant relationship between Post-Injury Play Time and our chosen predictor variables, and if the model derived from the first two years of data can predict a player's Post-Injury Play Time in the third season.

Variables

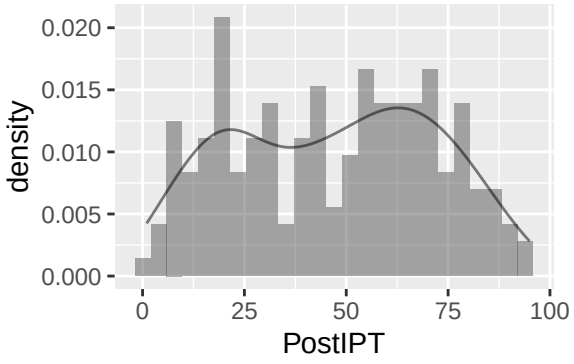
Our response variables is Post-Injury Play Time. Our explanatory variables that we are using are amtGamesMissed, PreIPT, WeekOfInjury, AgeAtConcussion, PreseasonInjury, and Position. PreseasonInjury and Position are qualitative variables and amtGamesMissed, PreIPT, WeekOfInjury, and AgeAtConcussion are quantitative variables.

Step 1: Univariate analysis - descriptive stats/plots for each variable

RESPONSE VARIABLE:

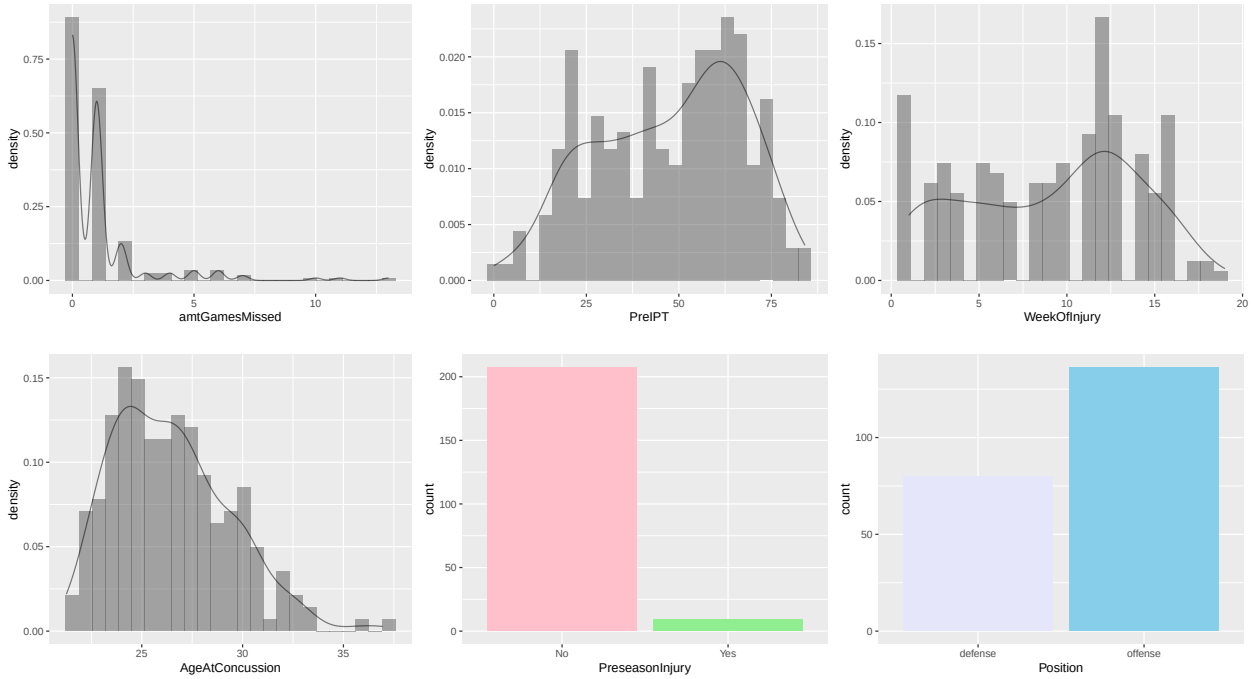
Post-Injury Play Time - Descriptive statistics and a frequency histogram with a density plot are shown below.

```
##  min Q1 median    Q3 max   mean     sd   n missing
##    1 25   48.5 67.25  95 46.859 24.753 184      32
```



EXPLANATORY VARIABLES:

amtGamesMissed, PreIPT, WeekOfInjury, and AgeAtConcussion are shown below as histograms. PreseasonInjury and Position are shown below as tallies.



Step 2: Compare variables in Bivariate Manner - Stats, Plots, and Compare

For step 2, we will first use GGpairs to compare each of our predictors (amtGamesMissed, PreIPT, WeekOfInjury, AgeAtConcussion, PreseasonInjury, and Position) with our response variable, PostIPT. This will help us compare each predictor with PostIPT. For two quantitative variable comparisons (amtGamesMissed, PreIPT, WeekOfInjury, and AgeAtConcussion), we will represent the relationship with a scatterplot. For one qualitative and one quantitative variable (PreseasonInjury and Position), we will use box plots. GGpairs gives us the correlation between two quantitative variables, and we will use this to identify correlation as well as any cases of multicollinearity. If there is indeed multicollinearity, we can use VIF > 5 to determine which variables are responsible and remove variables accordingly.

We will make the following comparisons in a bivariate manner:

Post-Injury Play-Time vs. Position.

We will represent this with a box plot, where the two categories are defense and offense.

Post-Injury Play-Time vs. Pre-Season Injury.

We will represent this with a box plot, in which the two categories are "No," the player didn't sustain an injury in the pre-season, and "Yes," the player did sustain an injury in the pre-season.

Post-Injury Play-Time vs. Post-Injury Play Time.

We will represent this with a scatterplot.

Post-Injury Play-Time vs. AgeAtConcussion.

We will represent this with a scatterplot.

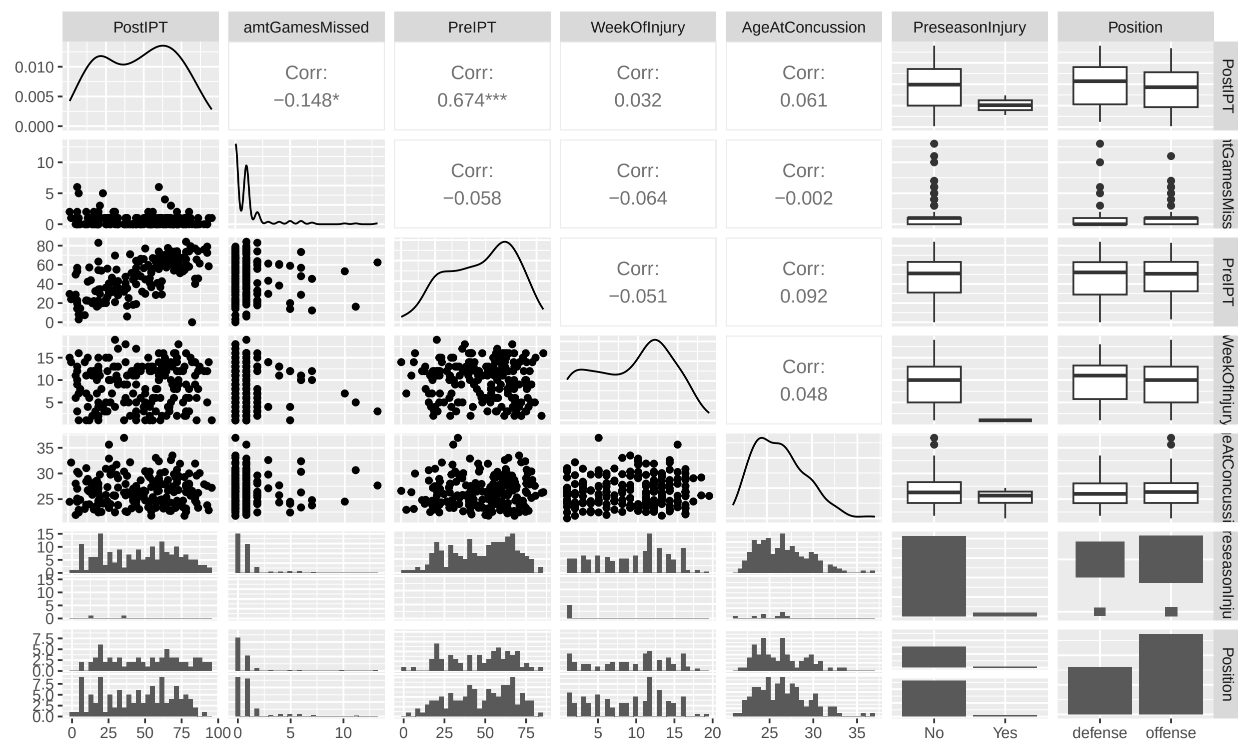
Post-Injury Play-Time vs. Week of Injury.

We will represent this with a scatterplot.

Post-Injury Play-Time vs. amtGamesMissed.

We will represent this with a scatterplot.

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Step 3: Model Building and Variable Selection - Choosing Variables, Checking/Evaluating Fit, and Final Model Criteria

For step 3, we will create a model with the variables remaining from the previous step (those that are non-multicollinear). We will assess the model's appropriateness using residuals vs. fitted plots and normal

qq-plots. If a variable does not meet the conditions, we will consider transforming the data (e.g., square root, log) and reassess. To evaluate the model fit, we will look for a high R-squared and low standard error.

Next, we will use the best subset method for variable selection in R to determine the most appropriate model for each number of predictors. We chose this automated method because it selects the best model for the given number of predictors. We will look for a low Mallow's CP and a high Adjusted R-squared. Based on the output, we will select the model with the fewest predictors that have a high Adjusted R-squared and low Mallow's CP.

We will also use AVplots to double-check if we missed any variables that provide significant information. Finally, with our chosen model, we will check the residuals vs. fitted plot, normal qq-plot, Adjusted R-squared, standard error, and p-values. If the conditions are not met, we will consider transforming the data and reassessing the conditions.

If we end up with multiple models that are similar in terms of Mallow's CP, Adjusted R-squared, standard error, and number of predictors, we will perform a nested F-test to determine if the more complex model is worth the extra predictors.