NBA Data Statistics Report for seasons 2021 - 2022

Victor Luque

Introduction

As of today, the NBA is one of the most prestigious basketball leagues in the world. The competition between teams is highly competitive. Apart from the economic power that most teams possess, the professionalism and skillfulness of players are what makes the NBA stand out from other leagues around the world. The report will cover statistics-based analysis which will focus on the areas of focus requested for the 2021-2022 regular season. The data provided in the file *NBA 2021 2022 Regular Season Data* was very useful as I did not have any issues carrying out my analyses to convey my findings to you in the most accurate way possible. The analyses were revised thoroughly to discern misleading and/or unprecise information. This report will give you a clear insight of the different evidence, differences, and significances that were gathered by performing multiple approaches of statistical analyses. The analysis strategies in which I used the data provided along with their respective interpretations for this report will be presented to you in the following paragraphs.

Analyses and interpretations:

NBA Players Usage Rate.

In order to state whether or not there was evidence that NBA players' usage rate was less than twenty percent of the time, I tested the statement put in for validation via the use of hypotheses testing for proportions. I needed to see which hypothesized statement came to be true. The p-value approach revealed clearly that the null hypothesis was the one to be validated and

consequently the alternative hypothesis was to be rejected. The results of my hypotheses testing concluded that there is not enough evidence that the usage rate of NBA players is less than twenty percent at a significance of 5 percent. This meant that even though some players are never used, most of them are moderately used as expected in a league that is recognized as one of the most competitive in the world.

Difference in Points Scored: Drafted vs Undrafted players.

t-*test: Two-sample Assuming Unequal Variances*

t-Test: Two-Sample Assuming Unequal Variances

	Drafted	*Undrafted*
Mean	9.984122563	4.405263158
Variance	57.10452374	13.67007839
Observations	359	95
Hypothesized Mean Difference	0	
df	315	
t Stat	10.135659	
P(T<=t) one-tail	2.21049E-21	
t Critical one-tail	1.649705334	
P(T<=t) two-tail	0.000	
t Critical two-tail	1.967523532	
Tstat	10.135659	
CV	1.967523532	
alpha	0.05	pvalue<alpha
pvalue	0.0000	

The findings of this analysis showed that there are differences between points scored by drafted and undrafted players. The statistical approach required for this analysis was a two-sample test assuming unequal variances. Firstly, I set up the null and alternative hypothesis, each stating if there is a difference (H0) or not (H1) in the number of points scored by drafted and undrafted players in the NBA 2021-2022 season. After performing the two-sample test, I could observe that the average amount of points scored by drafted players was 9.98. On the other hand,

undrafted players had a little less than half of the points scored standing at an average of 4.41 points per player. The difference in points between drafted and undrafted players was about 5 and a half points. This made sense because drafted players are likely to have more practice, opportunities, and commitment to score whereas undrafted players have fewer of the mentioned opportunities. By using the two-tailed test for unequal variances I examined that the test's p-value was 0.00; since the p-value was less than the significance level of five percent I could confirm that there is enough evidence between the average number of points scored by drafted vs undrafted players in the NBA 2021 – 2022 season.

Multiple regression to predict number of points scored by a player for the 2021-2022 season.

For this analysis I started by determining the explanatory and response variables to be used for the multiple regression. Upon revising the data set, I observed that a categorical variable was to be included to create a prediction model. In this data set this variable was a player's conference. After performing the regression, I could find that 67.50% of the variations in points scored by a player can be explained by the variations in games played, number of rebounds, number of assists, and conference they played in; after adjusting for the number of explanatory variables and sample size. While analyzing the ANOVA table from the regression and using the p-value approach I was able to observe that there is evidence of a linear relationship between number of points scored by a player and Games played, number of rebounds, number of assists, and conference taken together. (As a whole)

Multiple Regression ANOVA table

ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	*Alpha*
Regression	4	16315.0428	4078.7607	236.261991	0.0000	0.05
Residual	449	7751.409977	17.2637193			
Total	453	24066.45278				

The table of variable values on the regression performed, spotted a concerning finding when looking to detect which explanatory variable had the most impact when scoring points. Using the test for individual significance along with the p-value approach, I could examine that the categorical variable conference was not in alignment with the other variables as its p-value recorded .71 which is greater than the level of significance of 0.05. This inconvenience led to the assumption that creating a model with these variables was not yet a model that could best predict the number of points scored by a player.

Multiple Regression Table of Values

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.62599741	0.539244301	-1.16087905	0.24630791	-1.68575346	0.43375864
GP	0.2579206	0.058253671	4.42754231	0.0000	0.1434369	0.37240429
REB	0.96514667	0.090833526	10.6254453	0.0000	0.78663504	1.14365829
AST	1.90417215	0.110800658	17.1855672	0.0000	1.68641989	2.12192442
Conference	-0.14312761	0.390308687	-0.36670362	0.7140	-0.91018623	0.62393102

The next step following this frivolous difficulty was to exclude the categorical variable (conference) since the p-value was way too high and did not permit to formulate an optimal prediction model. After rerunning the regression, the data shown in the regression table had a positive outlook. I observed that the adjusted r-square value was .02% higher than the one calculated in the previous regression table. Using the confidence intervals from the regression, I was able to find the explanatory variable that held the most impact when scoring points along with the significance from all other variables that also had significance but were not as significant.

Explanatory variables impact by number of points; <u>most to least significant</u> at a 95% Confidence Level:

- For an increase of 1 assist (AST) the number of points scored by players on average increased by a minimum of 1.66 points and a maximum of 2.12 points, holding all other variables constant.
- For an increase of 1 Rebound (REB) the number of points scored by players on average increased by a minimum of 0.77 points and a maximum of 1.14 points, holding all other variables constant.
- For an increase of 1 game played (GP) there exists a 95% confidence that the number of points scored by players on average increased by a minimum of 0.14 points and a maximum of 0.37 points, holding all other variables constant.

Following these findings, I generated a model of the best fit for you to predict with ease the number of points scored by a player based on the given explanatory variables. The model of best fit and their respective legends are shown below:

Number of Points Scored: PTS.
Number of Games Played: GP.
Number of Rebounds: REB.
Number of Assists: AST.

$$PTS=-0.703+0.259*(GP)+0.965*(RB)+1.90*(AST)$$

For you to use this prediction model you need to input a number of games played, number of rebounds, and a number of assists. After doing so, you will be able to predict the number of points scored based on all the mentioned variables.

Conclusion and Recommendations

The overall results of my statistical analyses and interpretations were made with the utmost diligence for your clear understanding of the statistical data for the NBA 2021-2022 season. This report permitted me to discover that the amount of usage rate per player in the NBA was over 20 percent on average, it also allowed me to find forecasted differences when it comes to the scoring of points by an NBA player between drafted and undrafted players, I was also able to generate a best fit model to predict the number of points scored by an NBA player based on several explanatory variables. Although this data sufficed to carry out my analyses and interpretations to present this report to you, having more explanatory variables and adding more seasons in the data provided would have an increased report quality accuracy and prediction wise, this is because the more information we obtain to perform analyses, the more confidence there is in interpreting and conveying a more insightful report.