Analysis on James Harden's Game Performances and Nightlife Rank of Guest Game Cities

Andong Cai 1005035908

Dec 21st 2020

Title

Analysis on James Harden's Game Performances and Nightlife Rank of Guest Game Cities

Name of Author

Andong Cai

Date

December 21st, 2020

Abstract

In this report, I performed a linear regression analysis on game performance of James Harden, a pro NBA basketball player, and the guest city's nightlife rank to determine whether a relationship existed between the two variable. The linear regression model suggested that to a statistical significant level, there was not a relationship between James Harden's performance and the guest city's nightlife rank. In another word, Harden's nightlife did not influence his performance on court. Result of this analysis contradicted the general belief, by many NBA fans, that Harden's flucuating performance was due to his personal nightlife.

Keywords

NBA, James Harden, Correlation, Linear Regression, Performance, Nightclub, Observational Study

Introduction

The National Basketball Association, NBA, is an American basketball league that consists of 30 teams from different cities. It is one of the four major professional sports leagues in the United States and Canada, and it is considered as the top league for basketball in the world. Statistical analysis in sports is crutial because it can improve player performance, prevent injuries, increase revenue, and etc. Observational data of indivudual player are often drawn to make analysis on the player's performance. As one of the biggest stars in the NBA, James Harden is known for his exceptional basketball skill on court, as well as his obsession with nightclubs. Throughout the past few years, James Harden's performace has not been very consistent from games to games. Thus, many fans believe that Harden's performance on court are strongly affected by his nightlife.

In this research, I want to find out whether Harden's inconsistency in performace is dependent on his obession with nightclub. One way to find out whether there exists a correlation between the two variable is by performing a regression analysis on James Harden's performance and the guest citie's nightlife rank. Linear regression analysis is a linear approach to model the relationship between a response variable and one or more explanatory variables. (Wikipedia, 2020) Since I am analysing the relationship between Harden's performance and the guest city's nightlife rank, the explanatory variable would be guest city's nightlife rank, and the repsonse variable would be Harden's performance.

A total of four data sets will be drawn to analyze the relationship between James Harden's game performance and the guest city's nightlife rank. In the Data section under Methodology, I will describe the source of data that was used in the analysis as well as the cleaning process of data. Then, in the Result section, I will present the result of linear regression analysis on the data. In the Discussion section, I will make intepretations of the result, and further discuss about weaknesses of the analysis.

Methodology

Data

```
## Warning: Missing column names filled in: 'X6' [6], 'X8' [8]

## Warning: Missing column names filled in: 'X6' [6], 'X8' [8]

## Warning: Missing column names filled in: 'X6' [6], 'X8' [8]
```

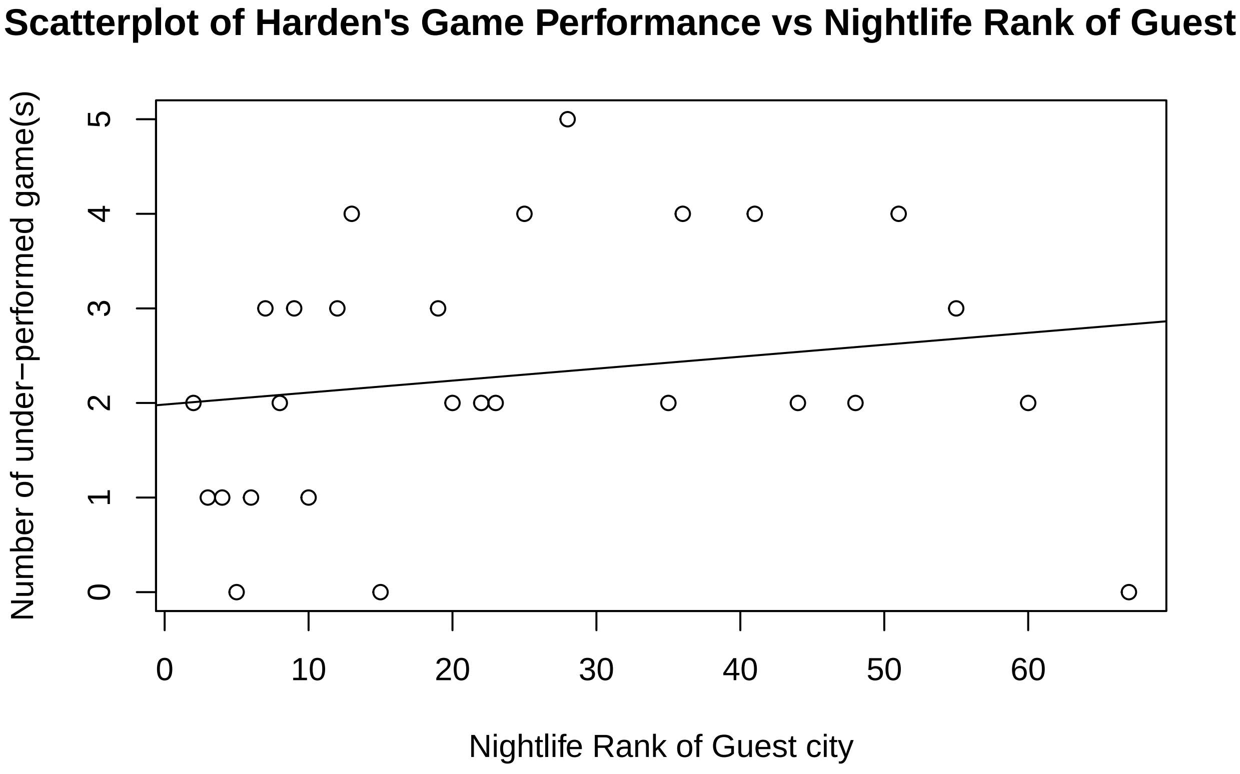
I extracted James Harden's performance statistics from Basketball Reference, a professional website that provided statistics for major basketball leagues(Harden, 2020). For the purpose of this analysis, I decided to use game statistics of James Harden from seaon 2016-2017, 2017-2018, and 2018-2019. The population was all of Harden's game statistics, and sample data is the three chosen season's game statistics. There were several reasons for choosing these three specific seasons. First of all, by choosing more than one season's game statistics, I could get more samples, thus generating a more accurate mean value for James Harden's performance. In addition, these three season provided up-to-date data, and during 2016-2019, Harden's team structure was very consistent, meaning no major player trade in the team. I purposefully did not include data of 2019-2020 season becauce covid-19 pandemic caused NBA to suspend the season starting on March 11, 2020(ESPN, 2020). When the season restarted, NBA had all teams move to Orlando for games, practices, and housing. This meant that there was no home game or guest game for a majority time of 2019-2020 season, thus it was reasonable to exclude this data from my analysis. I collected data for nightclub score of each guest city on WalletHub(McCann, 2020), which provided a "Nightlife and Party" rank for each city in the States.

To start processing the data, I first downloaded James Harden's performance statistics from Basketball Reference for 2016-2017, 2017-2018, 2018-2019 seasons. Then, I read the data into my R studio. There were 30 variables in the data, however, for the purpose of this analysis, I only used variable Opp, which stands for opponent of the game, and GmSc, which stand for game statistics. In variable opp, it list the opponent team's abbreviation. The variable GmSc is first developed by John Hollinger to give a total value for the player's performance in a basketball game(NBA, 2020). To cleaning up the data, I first excluded all games where James Harden is absent or did not play. Then, I excluded all home games from the data. There were variables in raw data, such as field goal, points, assistance that also represents a player's performance. However, game score was a complete assessment on all of these variables. Thus, it was better to use game score rather than a combination of all the other variables. I calculated the average game score of each season, and I filtered all the games where Harden's game score is below season average.

Model

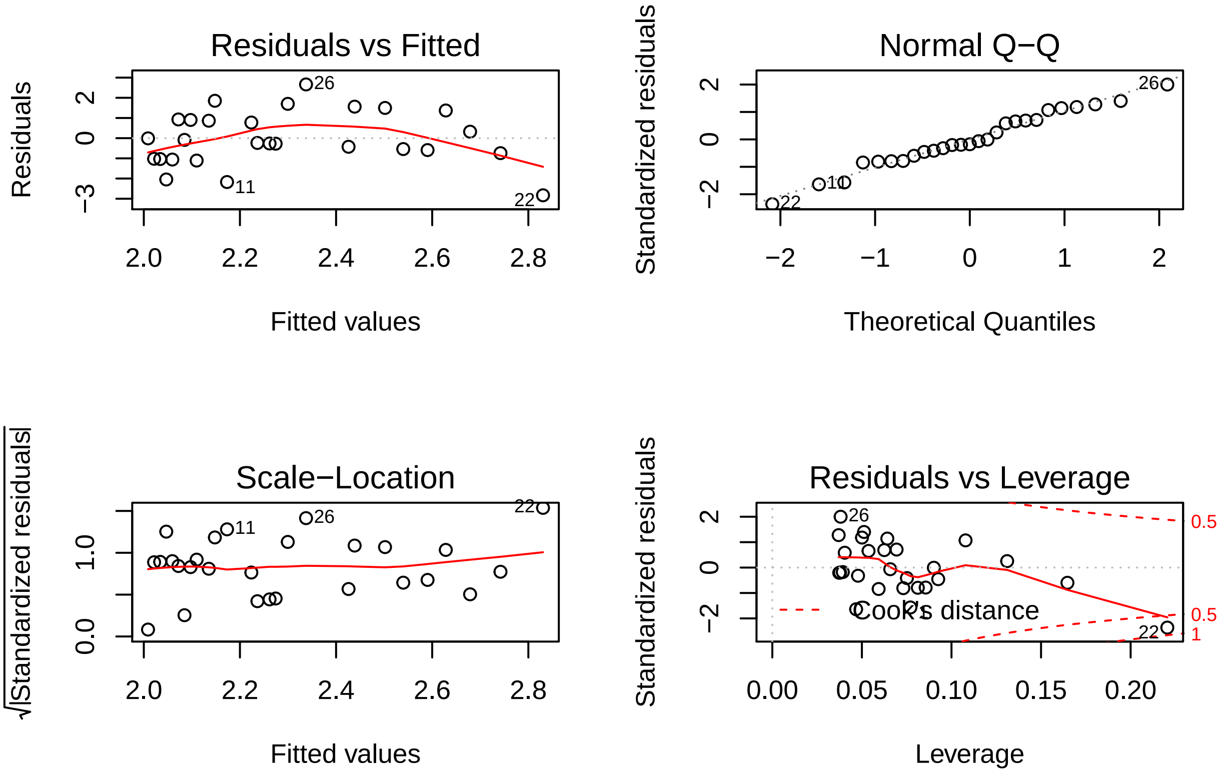
In this analysis, my goal to find out the relationship between James Harden's game performance and the guest city's nightlife rank using R studio. The explanatory variable in the analysis is guest city's nightlife rank, and the response variable is the number of under performed game(s) of Harden. To find whether a relationship exists between the two variable, I plotted a scatterplot, and then fit a simple linear regression line. The model consists of one predictor, guest city's nightlife rank. This variable represent the rank of guest city's nightlife, and it is calculated by entertainment & recreation, nightlife & parties, and cost. For example, Orlando is ranked the 2nd in the States, and it has the best overall rank among all cities that have a NBA team, whereas Phoenix is ranked the 67th in the States, and it has the worst rank among all NBA cities. I choose the nightlife rank of city as explanatory variable because it is a good indicator of how much Harden is influenced by the guest city. The reason to choose game score as response variable is because game score provide a complete analysis on the player's performance in each game. The complete model is given by, $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} * x_{nightlife-rank}$. After I plug in the value of β_0 and β_1, I'll get the follwing model: $\widehat{y} = 1.98 + 0.01264 * x_{nightlife-rank}$

Results



In this seciton, I will be focusing on the scatterplot of Harden's game performance vs. nighlife rank of guest city an also the linear regression model discussed in the previous section. As you can see from the scatterplot, all points are scattered, and there seems not to be an linear relationship. Based on the model generated from the scatterplot, $\widehat{\beta_0}$, the estimated intercept term, is 1.98, and its p-value is 9.92e-05. The estimated coefficient for nightlife rank, $\widehat{\beta_1}$, is 0.013, and its p-value is 0.367. These resaults are based on the analysis of linear regression model, and the summary statistics of the model can be found in the Appendix section. Below is a table of estimated value and p-value of each coefficient.

Coefficient	Estimate	p-value
Intercept	1.98	9.92e-05
Nightlife rank	0.013	0.367



Above are the regression diagnostic plots. From the Residuals vs. Fitted value plot, it indicates that our model shows non-linear relationships. In the normal Q-Q plot, we can see that residuals are normally distributed. Moving on to Scale-location plot, we can see a horizontal line with equally spread point, which indicates equal variance. In the last plot, residuals vs leverage, we can see that data number 22 is an influential point to our model.

Discussion

Summary

In this analysis, I used four datasets, James Harden's 2016-2017, 2018-2018, 2018-2019 game statitstics and NBA guest cities' nightlife rank to build a simple linear regression model. Then I used the model to analyze relationship between Harden's game performance and the guest city's nightlife. I retrieved my data from Basketball Reference and Wallethub. Using R code, I first cleaned Harden's game statistics data, and then plotted the scatterplot with nightlife rank as the explanatory variable and number of sub performance games as the response variable. In choosing a measurement for Harden's performance, I did not choose to use individual statistics, like field goals, points, assists etc. Instead, I used game score, which is an overall evaluation of a player's performance. This may create potential bias on the model, and may eventually influence the result. After plotting the data, I fitted an linear regression model to determine the coeeficient for intercept and night life rank.

The linear regression analysis showed that for each increase in rank of guest city's nightlife, James Harden's under-performed game would go up by 0.013. However, the p-value of the coeeficient is 0.0367, which was not statistically significant. Thus, we can not make a conclusion on wheather James Harden's game performance has a relationship with the guest city's nightlife rank. The intercept coefficient term had a p-value of 9.92e-05, which is less than 0.05. But, the intercept term was not meaningful in the context.

Weaknesses

There are varias factors that can influence a basketball player's performance on court, and these factors are not accounted in my model. For example, time difference between different city, diet, shoes, and player's mood on court can all inflence a player's performance in a very significant way. I did not include these variables in my model because there was not a numerical measurement to measure these factors. However, if these factors can be included in my model, I would expect to see a very big change in result. Additionally, the rank by nightlife of each city is not the same in different websites. Since the rank can be very subjective, the same game statistics data can show different result if I choose to use data of nightlife rank from another source.

Next Step

To improve my model, I can do a tranformation on my dataset to improve the linearity in residual vs fitted value plot. I can also take out influential points found in Residual vs leverage plot to improve my model.

References

ESPN. "NBA Suspends Season Due To Coronavirus". ESPN.Com, 2020, https://www.espn.com/nba/story/_/id/28887560/nba-suspends-season-further-notice-player-tests-positive-coronavirus. Accessed 21 Dec 2020.

Harden, James. "James Harden Stats | Basketball-Reference.Com". Basketball-Reference.Com, 2020, https://www.basketball-reference.com/players/h/hardeja01.html. Accessed 21 Dec 2020.

McCann, Adam. "Most Fun Cities In America". Wallethub, 2020, https://wallethub.com/edu/most-fun-cities-in-the-us/23455. Accessed 22 Dec 2020.

NBA. "Game Score". Nba.Com, 2020, https://www.nba.com/resources/static/team/v2/thunder/statlab-gamescore-191201.pdf. Accessed 21 Dec 2020.

Wikipedia. "Linear Regression". En.Wikipedia.Org, 2020, https://en.wikipedia.org/wiki/Linear_regression. Accessed 21 Dec 2020

Appendix

```
##
## Call:
## lm(formula = number_of_underperformance ~ rank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8306 -0.8818 -0.2364  0.9153  2.6625
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.98352    0.42911   4.622 9.92e-05 ***
## rank         0.01264    0.01376   0.919    0.367
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.358 on 25 degrees of freedom
## Multiple R-squared:  0.03267,    Adjusted R-squared:  -0.006026
## F-statistic: 0.8443 on 1 and 25 DF,  p-value: 0.367
```