511_project_report

Neil Pareddy, Shivam Zala, Yu Kang, Zaire Wade, Zhonghao Xue

12/8/2021

Introduction

There are certain debates you can count on every NBA season. One of them is a week-long debate in late January surrounding which players should and shouldn't have been named All-Stars. Both fans and sports analysts can't agree on who the top-12 players in each conference are, so how are we even supposed to figure out who should be on the 75th Anniversary Team?

The NBA 75th Anniversary Team was selected by a blue-ribbon panel of current and former NBA players, coaches, general managers, and team and league executives, WNBA legends, sportswriters, and broadcasters. Voters were asked to select the 75 Greatest Players in NBA History without regard to position. Panelists did not rank their selections. Current and former players were not allowed to vote for themselves. Note, Dave DeBusschere is the only player not included in this project dataset.

In this project, we attempt to settle this debate statistically; by comparing the statistics of players listed on the 75th Anniversary to those who were "snubbed'.

Those labeled as "snubbed" mean that players did not make the anniversary list; however, public sentiment suggests that they should've. A list of snubbed player names is generated from public opinion, in which suggested names from NBA reporters and analysts are used. Tony Parker, Dwight Howard, and Alex English are just some of the surprising names that were snubbed from the list.Why is that? What statistical trends favor those who made the list? How important is total points scored in a player's career, when deciding whether the player deserves recognition on the 75th Anniversary. These are a few questions that will be explored in this project's analysis.

Statistical Methods

1. About the data

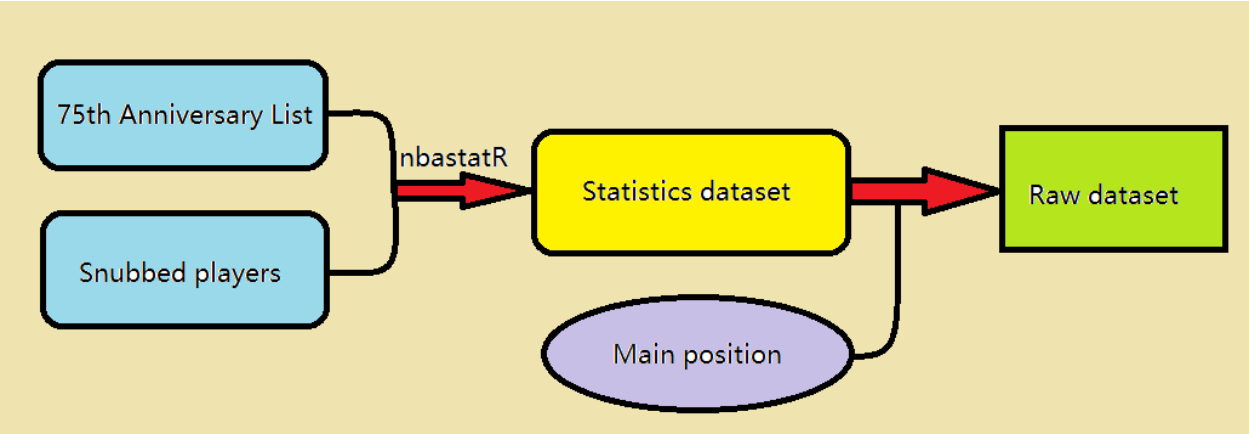
Starting from names of the 75th Anniversary List players and the suggested-but-snubbed players, in order to explore their statistical differences, the dataset we need would necessarily provide statistical data of NBA players such as total score, rebound, steal, assist, etc. Thus, we have picked the library of nbastatR since it could provide access to data sources include, but not limited to: NBA Stats API, Basketball Insiders, Basketball-Reference, HoopsHype, and RealGM. The player stats are acquired through functions such as "players_careers" and "get_nba_players_ids" to collect career regular-season statistics for all players involved in this project.

Thus, using the data retrieved from "nbastatR" package, we were able to assemble a dataframe of players either in list or snubbed from the list. However, the dataset contains no information of a player's exact position, so there's another question to be asked: as a player's main position determines their certain tasks in matches and could greatly influence their pattern of statistics, how do we compare players appropriately to make our result concrete for players from various positions? Well, in order to make viable analysis, we

decided to add a categorical variable called "position" to manage such information, and to manually impute the main position for each player (snubbed/listed) from google search; thanks to the relatively small size of the name list we were able to do such manual imputation, and our data is now assembled into a single dataframe that contains a player's name, their statistics, their status of snubbed/listed, and their main position.

Filter

	Name	idTeam	idPlayer	slugSeasonType	gp	gs	fgmTotals	fgaTotals	pctFG	fg3mTotals	fg3
1	Adrian Dantley	0	76504	RS	955	900	8169	15121	0.540242	7	
2	Allen Iverson	0	947	RS	914	901	8467	19906	0.425349	1059	
3	Anthony Davis	0	203076	RS	587	582	5228	10165	0.514313	268	
4	Artis Gilmore	0	600014	RS	909	476	5732	9570	0.598955	1	
5	Bernard King	0	77264	RS	874	795	7830	15109	0.518234	23	
6	Bill Russell	0	78049	RS	963	NA	5687	12930	0.439829	NA	
7	Bill Sharman	0	78126	RS	711	NA	4761	11168	0.426307	NA	
8	Bill Walton	0	78450	RS	468	117	2552	4900	0.520816	0	
9	Billy Cunningham	0	76487	RS	654	NA	5116	11467	0.446149	NA	
10	Bob Cousy	0	600003	RS	924	NA	6168	16468	0.374544	NA	
11	Bob Lanier	0	600005	RS	959	179	7761	15092	0.514245	2	
12	Bob McAdoo	0	77498	RS	852	1	7420	14751	0.503016	3	
13	Bob Pettit	0	77847	RS	792	NA	7349	16872	0.435573	NA	
14	Carmelo Anthony	0	2546	RS	1215	1120	9916	22183	0.447008	1645	
15	Charles Barkley	0	787	RS	1073	1012	8435	15605	0.540531	538	
16	Chauncey Billups	0	1497	RS	1043	937	4738	11413	0.415140	1830	



2. Cleaning the data

Then, in order to make the dataset suitable for further analysis, the procedures of data cleaning must be applied.

First of all, we checked the existence of missing values and found that the features are either complete or with a big ratio of missing values, usually higher than 10% and reaching a level that simple imputation seems not to guarantee the integrity of information. Thus, while such greatly incomplete columns might be due to the library itself, we had no choice but to remove them from our dataset. We also removed the redundant features of player identity.

Filter

	Name	idTeam	idPlayer	slugSeasonType	gp	fgmTotals	fgaTotals	pctFG	pctFT	minutesTotals	ftm
1	Adrian Dantley	0	76504	RS	955	8169	15121	0.540242	0.818105	34151	
2	Allen Iverson	0	947	RS	914	8467	19906	0.425349	0.780484	37578	
3	Anthony Davis	0	203076	RS	587	5228	10165	0.514313	0.796116	20223	
4	Artis Gilmore	0	600014	RS	909	5732	9570	0.598955	0.713245	29685	
5	Bernard King	0	77264	RS	874	7830	15109	0.518234	0.729610	29417	
6	Bill Russell	0	78049	RS	963	5687	12930	0.439829	0.560741	40726	
7	Bill Sharman	0	78126	RS	711	4761	11168	0.426307	0.883113	21793	
8	Bill Walton	0	78450	RS	468	2552	4900	0.520816	0.660130	13250	
9	Billy Cunningham	0	76487	RS	654	5116	11467	0.446149	0.719525	22406	
10	Bob Cousy	0	600003	RS	924	6168	16468	0.374544	0.803335	30165	
11	Bob Lanier	0	600005	RS	959	7761	15092	0.514245	0.766570	32103	
12	Bob McAdoo	0	77498	RS	852	7420	14751	0.503016	0.754255	28327	
13	Bob Pettit	0	77847	RS	792	7349	16872	0.435573	0.761423	30690	
14	Carmelo Anthony	0	2546	RS	1215	9916	22183	0.447008	0.813780	42404	
15	Charles Barkley	0	787	RS	1073	8435	15605	0.540531	0.734582	39331	
16	Chauncey Billups	0	1497	RS	1043	4738	11413	0.415140	0.894014	33009	
17	Chris Bosh	0	2547	RS	893	6209	12581	0.493521	0.798783	31935	
18	Chris Paul	0	101108	RS	1113	7151	15140	0.472324	0.872117	38500	
19	Clyde Drexler	0	17	RS	1086	8335	17673	0.471623	0.787990	37538	

Secondly, when we check the distributions of player statistics, we found an extreme outlier with gp=7 (only played 7 games) while other players all had at least several hundred of games. This outlier, with a player name of Patrick Ewing, seems to coincide with another normal entrie named Patrick Ewing, so we concluded the outlier might come from weird statistics of the library, and decided to remove it from the dataset.

Finally, in the original columns of our dataset there were 4 variables fgmTotals, fgaTotals, ftmTotals and ftaTotals mainly describing the total number of shootings each player has attempted and successfully made, as either field goals or free throws. Unfortunately, we already had variables pctFG and pctFT describing the percentage of shootings, exactly as the ratio of the previous 4 variables. (pctFG=fgmTotals/fgaTotals, pctFT=ftmTotals/ftaTotals) Plus, there's already a variable gp describing the number of games each player has had, and by common knowledge gp is also highly correlated with the 4 Total shooting features. Thus, in order to preserve the meaningfulness of our models like multiple linear regression, we had to discard either the 4 Totals shooting features or gp+the 2 ratio variables, and we decided to drop the previous set.

Now, our cleaned dataset has totally 91 observations and 11 columns(one for player names, one for response variable Status, and the rest 9 as predictors). Our goal is to use this dataset to figure out the statistical differences between snubbed players and listed players.

Filter

	Name	gp	pctFG	pctFT	minutesTotals	trebTotals	astTotals	pfTotals	ptsTotals	Status	Position
1	Adrian Dantley	955	0.540242	0.818105	34151	5455	2830	2550	23177	0	SF
2	Allen Iverson	914	0.425349	0.780484	37578	3394	5624	1777	24368	1	PG
3	Anthony Davis	587	0.514313	0.796116	20223	5998	1363	1375	14024	1	PF
4	Artis Gilmore	909	0.598955	0.713245	29685	9161	1777	2986	15579	0	C
5	Bernard King	874	0.518234	0.729610	29417	5060	2863	2885	19655	0	SF
6	Bill Russell	963	0.439829	0.560741	40726	21620	4100	2592	14522	1	C
7	Bill Sharman	711	0.426307	0.883113	21793	2779	2101	1925	12665	1	SG
8	Bill Walton	468	0.520816	0.660130	13250	4923	1590	1298	6215	1	C
9	Billy Cunningham	654	0.446149	0.719525	22406	6638	2625	2431	13626	1	SF
10	Bob Cousy	924	0.374544	0.803335	30165	4786	6955	2242	16960	1	PG
11	Bob Lanier	959	0.514245	0.766570	32103	9698	3007	3048	19248	0	C
12	Bob McAdoo	852	0.503016	0.754255	28327	8048	1951	2726	18787	1	C
13	Bob Pettit	792	0.435573	0.761423	30690	12849	2369	2529	20880	1	PF
14	Carmelo Anthony	1215	0.447008	0.813780	42404	7610	3375	3433	27713	1	SF
15	Charles Barkley	1073	0.540531	0.734582	39331	12546	4215	3287	23757	1	PF
16	Chauncey Billups	1043	0.415140	0.894014	33009	2992	5636	2169	15802	0	PG

3. Method Pipeline

First of all, we did a bunch of EDAs on the dataset exploring the distributions of statistics for listed players and snubbed players.

Then, as the EDAs seemed to suggest different performances of features, we decided to run a 2-step pipeline:

1. Run regression on Status and position to check the effect of different positions on the deciding threshold of 75th anniversary list.
2. Then, we would perform a series of methods and hypothesis tests on the conditional distributions of players based on their positions. The certain tests would be listed in following terms.
3. The first method we would use is multiple linear regression. While the response variable is the binary Status of snubbed or not, we believe the significance levels and coefficient signs for variables could reveal a bit of the relationship between these variables and the choice of the list. While we utilized the method of backward selection, we started from including al the features and repeatedly removed features with highest p-value (that means, the least predictive one), until the rest variables have at most 1 insignificant variable at 20% level. Then, we picked the 3 variables of the least p-value for each position and interpreted the result through both significance level and sign of coefficient(positive means listed players generally have higher stat on the predictor, and negative stands for snubbed players).
4. Next, we performed permutation tests to check the distribution of our response variable Status with respect to each of other predictors at each level of position.

Null hypothesis h0: The Status variable is distributed evenly with respect to a pair of position/predictor.

Alternative hypothesis ha: The Status variable is not evenly distributed with respect to a pair of position/predictor, so such predictor could be important to predict the snubbed/listed status for players in the paired position.

The mechanism behind this method is that, if our null hypothesis were true, then permuting the Status column would not make a significant shift on the core statistic. Thus, by perform many permutations and recording the differences in means of predictor, we could get the null hypothesis distribution and see the probability that our observed difference in means between listed players and snubbed players, from which we could infer the comparison between performances of listed players and snubbed players.

5.Our last method to use is the 2 sampled T-test, which also checks the difference in means of our predictors between listed and snubbed players. We decided to run a 2-stage T-test for each pair of position/predictor; while the first stage encodes a 2-sided test to see whether the difference in means is non-zero, the second stage of one-sided test would be applied if the first stage test receives a significant result, and the second stage is used to determine the exact significance level of certain pair of position/predictor.

Null hypothesis h0: There's no significant difference in means of certain predictor for players in certain position with different Status (snubbed/listed).

Alternative hypothesis ha_1: There's significant difference in means.

Alternative hypothesis ha_2: The difference in means is greater(less) than 0. (Based on the observed difference in means)

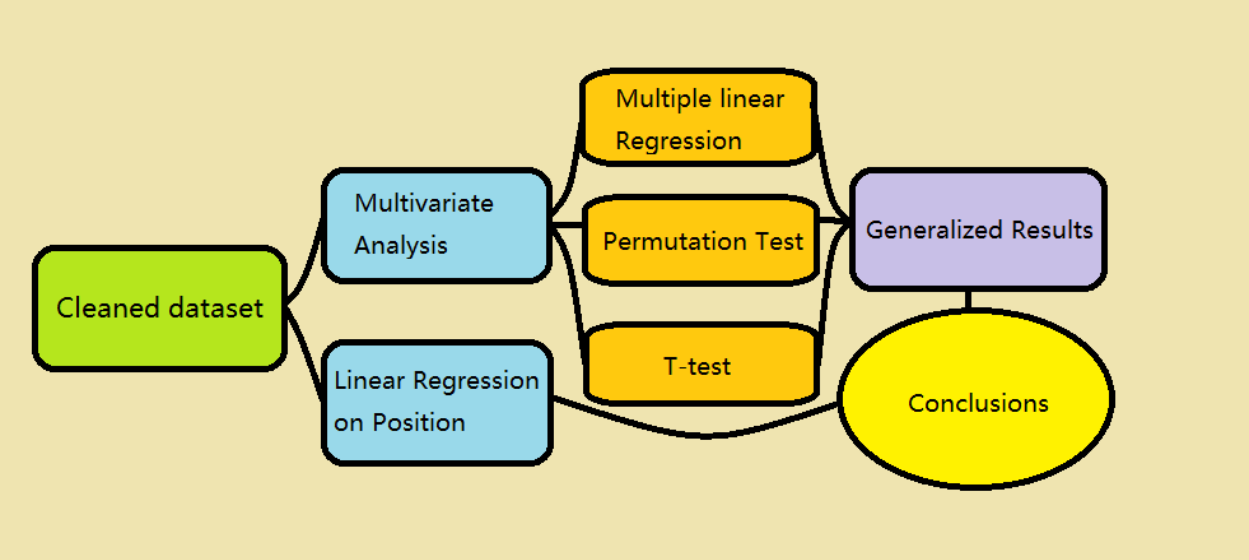
4. Generalization of the Results

Till now we have applied totally 3 different methods to each of the position/predictor pair, and in order to extract useful conclusions from it we decided to let the methods "vote" for themselves:

If at least 2 of the 3 models claim that listed players have a significantly higher level for certain certain and position pair, then the vote suggests that such predictor could be important when deciding whether players from the corresponding position should be accepted into the list.

Or, if at least 2 of the 3 models claim that snubbed players have a significantly higher level for certain certain and position pair, then the vote suggests that such predictor should be important when deciding whether players from the corresponding position should be accepted into the list, but was actually underestimated by the true list.

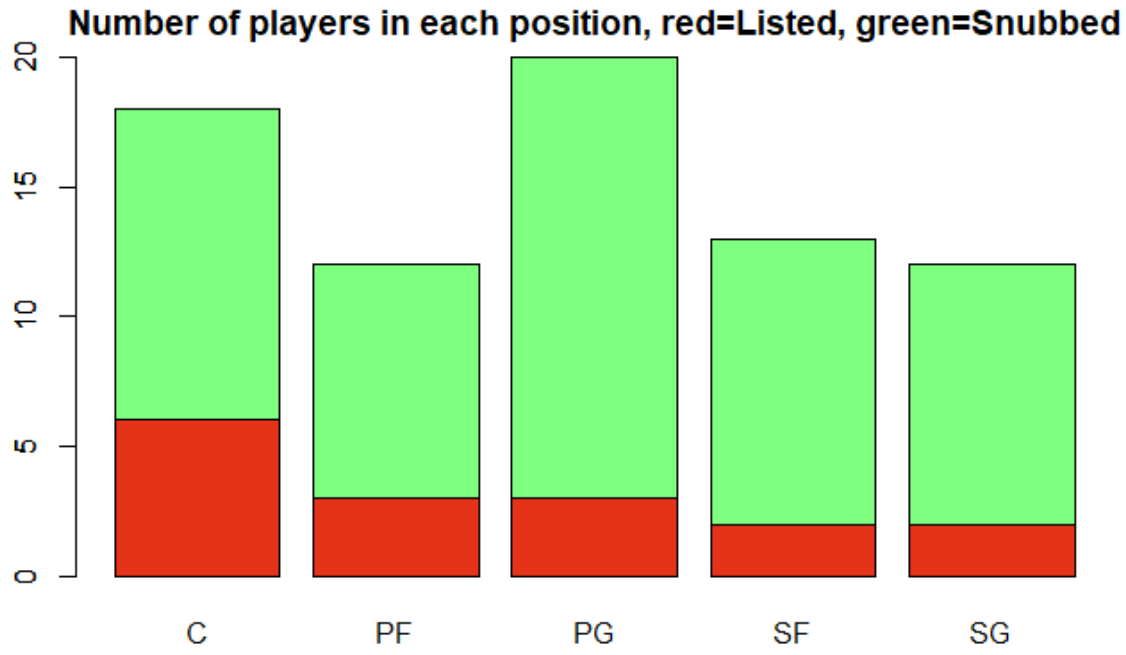
By such criterion we could conclude the possible main preferences for both the actual list and the suggestion makers of snubbed players, and by comparing these preferences to the actual tasks for each position we could derive our conclusions on the convincing power of these suggestion makers and possible reasons behind such inferences.



Results

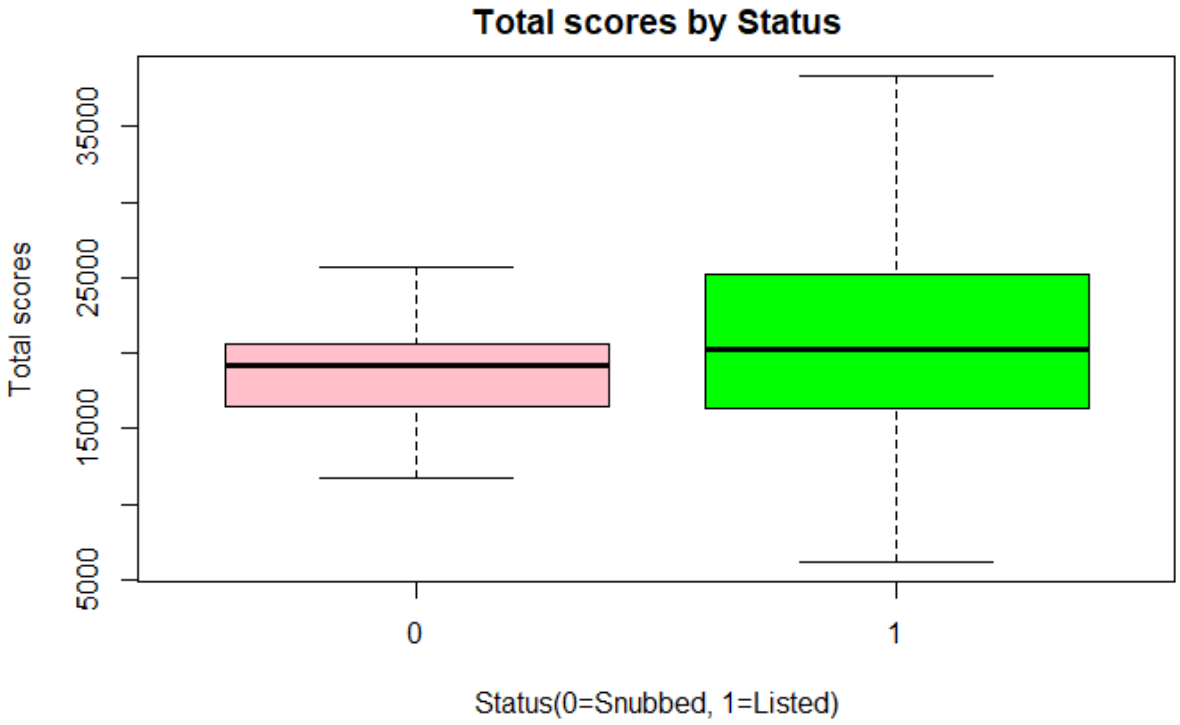
1. Exploratory Data Analysis (EDA)

(1) The number of players in each position



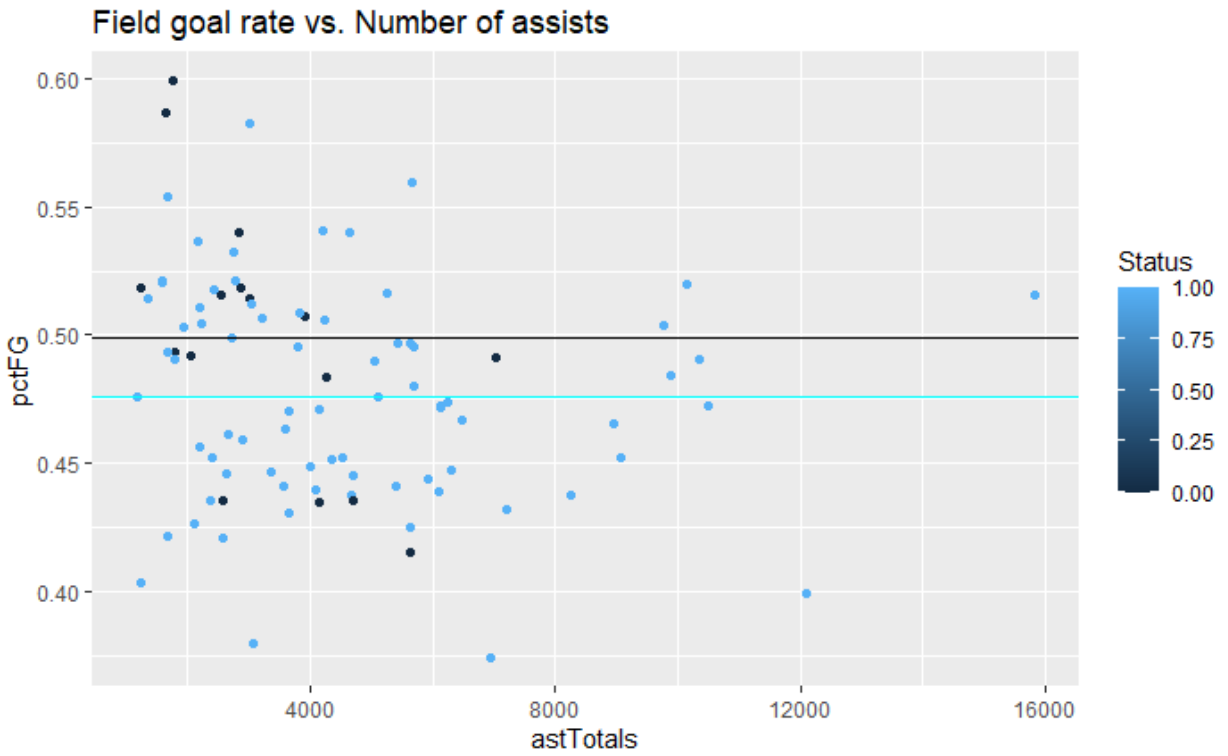
We observed that both the absolute and relative size of snubbed players are small, roughly 1/6~1/3 of that for listed players in each of the 5 positions, so the analysis results might be a bit unstable. Some may suggest the technique of oversampling on the snubbed players, but it might be not good to do that, as we are not sure whether our sample of snubbed players represents the overall level of all possible suggestions. Thus, we decide to keep the dataset as it is.

(2) Boxplots of total scores (ptsTotals) for listed players and snubbed players



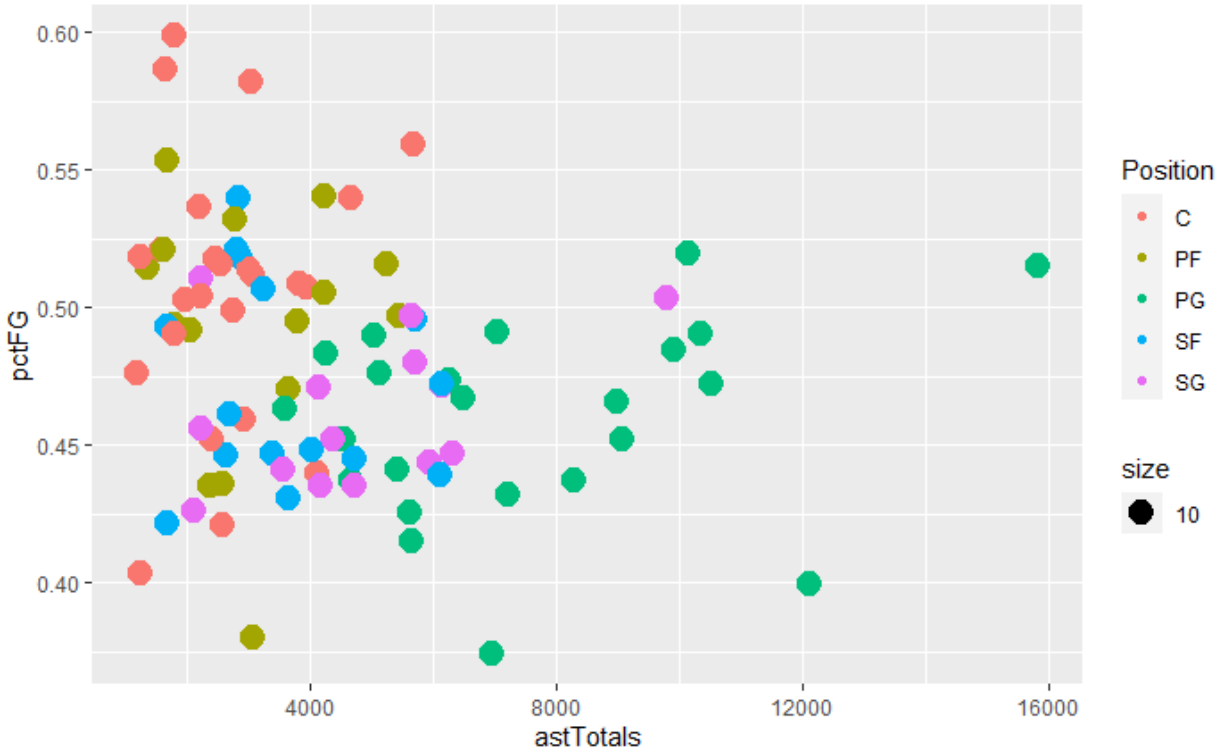
We observed that that while listed players have a generally wider distribution on total score, they have similar median of total score to that of snubbed players, so it might not be easy to extract useful inferences just by this plot, as the wider range of listed players comes possibly from a larger number of players.

(3) Scatter plot of field goal rate against number of assists



From this plot we can see that while listed players seem to have a wider distribution of number of assists and many players well at assistance, the observed snubbed players seem to have a higher average in field goal rates, but such relationships would need further validation to be properly argued.

(4) Relative advantages of each position



From plot we could see that while Center players generally have the highest level of field goal accuracy, Point Guard players seem to focus on teamwork instead of being a scorer, while the rest 3 positions could not be separated well from the current stage of data we obtained. This visualization might be helpful in the analysis stage where we would judge the fairness of suggestions of snubbed players, position-wise.

2. Regular Analysis

###(1) (Multiple) linear regression

In this project, we have utilized multiple linear regression model, permutation test and T-test to determine the possible factors deciding NBA basketball players for the 75th Anniversary List, and our first stage of analysis lands on whether players from a certain position are more likely to be selected for the list. The reason we set the starting point here is that, as patterns for player's statistics vary greatly depending on their position, if the selection of listed players has a preference over certain aspects(factors) of the statistics, such preference could possibly get expressed in the distribution of positions. While one of our EDAs explored such distribution for listed and snubbed players, a validation by proper model would be necessary.

For this stage we used the linear regression model, starting with the null hypothesis that there is no relationship between the position player is playing and the status that the player is listed or not. We ran the model for each level of the Position factor, and the R results were like:

```
Call:
lm(formula = Status ~ Position == "SG", data = nba)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8571  0.1429  0.1818  0.1818  0.1818

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.81818    0.04384  18.665   <2e-16 ***
Position == "SG"TRUE  0.03896    0.11176   0.349    0.728
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3847 on 89 degrees of freedom
Multiple R-squared:  0.001364,  Adjusted R-squared:  -0.009857
F-statistic: 0.1215 on 1 and 89 DF,  p-value: 0.7282
```

While based on the test statistics, since the p-value is always greater than 0.05, we fail to reject the null hypothesis at 5% level, while the R-square statistic were always too low to support the significance of the Position levels. Thus, we conclude that NBA basketball players from different positions all have a similar probability to be selected into the top 75 list, and for significant inferences we need to look into the detailed statistics for those players.

However, as our last EDA suggested that there could be different patterns of statistics for players from different positions, the criterion of selection for the 75th Anniversary List is likely to vary with respect to a player's position, and we need to utilize the idea of multivariate analysis to look at conditional distributions based on different levels of positions, in order to arrive at more robust results. So, as the features could have distinct level of importance, we divide the analysis into position/feature pairs; in this way, while our general form of null hypothesis states that the player listed or snubbed has a similar performance on a certain statistics, we would explore the potential of our alternative hypothesis that listed/snubbed players from the given position have a significantly better performance on the certain statistics.

Now, for the first step of our position-wise exploration, we also use the linear regression model to research on the NBA statistics on both the listed players and the snubbed players, but this time we would use the multiple linear regression model to explore the possibilities of various features in the dataset. An obvious advantage of multiple linear regression model is that we could explain a higher ratio of variance in the response variable Status, and for more accurate interpretations we have utilized the idea of backward selection, starting from the whole collection of predictors and repeatedly removing the predictor with highest p-value(that is, least predictive), until the remaining predictors meet our need for a solid interpretation. Then, in order to control the number of significant arguments, we would pick the 3 features with least p-value(still need to be at least significant in 10% level), marking them as important for list selection/suggestion and record their significance levels.

Here we used a stopping criteria of (at most 1 predictor with p-value > 0.20) based on our practical tuning of the models, and the example below could help illustrate how we arrived at the regression results:

Choosing the Small Forward position subset, we started by the whole model:

```
Call:
lm(formula = Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals +
    astTotals + pfTotals + ptsTotals, data = SF)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23215 -0.06935 -0.01154  0.05298  0.36309

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.720e+00  2.638e+00   1.031   0.3421
gp             5.575e-03  3.428e-03   1.626   0.1550
pctFG         -6.800e+00  2.544e+00  -2.673   0.0369 *
pctFT          1.408e+00  1.832e+00   0.769   0.4713
minutesTotals -1.418e-04  1.335e-04  -1.062   0.3291
trebTotals     1.363e-04  8.614e-05   1.582   0.1647
astTotals      8.688e-05  1.388e-04   0.626   0.5544
pfTotals      -3.822e-04  1.822e-04  -2.097   0.0808 .
ptsTotals     -2.689e-05  6.436e-05  -0.418   0.6907
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.225 on 6 degrees of freedom
Multiple R-squared:  0.8247,    Adjusted R-squared:  0.591
F-statistic: 3.529 on 8 and 6 DF,  p-value: 0.07089
```

The R-squared is fairly high, but the p-values suggest that our predictors are generally insignificant, and the F-test is even insignificant at 5% level, illustrating that our model still need to be improved, so we perform backward selection to remove the feature ptsTotals:

```
Call:
lm(formula = Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals +
    astTotals + pfTotals, data = SF)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27374 -0.04664 -0.02582  0.06432  0.33880

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.155e+00  2.276e+00   1.386   0.2082
gp             6.402e-03  2.627e-03   2.437   0.0449 *
pctFG         -7.276e+00  2.135e+00  -3.408   0.0113 *
pctFT          1.070e+00  1.543e+00   0.693   0.5106
minutesTotals -1.839e-04  8.231e-05  -2.234   0.0606 .
trebTotals     1.345e-04  8.080e-05   1.665   0.1399
astTotals      1.382e-04  6.067e-05   2.278   0.0568 .
pfTotals      -4.000e-04  1.664e-04  -2.404   0.0472 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2113 on 7 degrees of freedom
Multiple R-squared:  0.8196,    Adjusted R-squared:  0.6393
F-statistic: 4.544 on 7 and 7 DF,  p-value: 0.03189
```

This time, while the Multiple R-squared only experienced a minimal loss, the p-values of predictors are

generally smaller than the first model, the F-statistic becomes significant at 5% level, and the adjusted R-squared even has an increment, meaning that our model has improved by removing the predictor ptsTotals.

But this is still not enough for our criterion, and pctFT still has a high p-value of 0.51, so we remov

```
Call:
lm(formula = Status ~ gp + pctFG + minutesTotals + trebTotals +
    astTotals + pfTotals, data = SF)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29324 -0.07662 -0.04086  0.10540  0.32926

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.533e+00  1.072e+00   4.228  0.00288 **
gp             5.298e-03  2.020e-03   2.623  0.03052 *
pctFG         -7.933e+00  1.850e+00  -4.287  0.00266 **
minutesTotals -1.472e-04  6.095e-05  -2.415  0.04218 *
trebTotals     1.008e-04  6.239e-05   1.616  0.14486
astTotals      1.367e-04  5.863e-05   2.332  0.04800 *
pfTotals      -4.614e-04  1.362e-04  -3.388  0.00952 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 8 degrees of freedom
Multiple R-squared:  0.8072,    Adjusted R-squared:  0.6627
F-statistic: 5.584 on 6 and 8 DF,  p-value: 0.01483
```

Now, we could observe that the predictors are generally significant at their distinct levels, meeting our criterion. So, we could starting analyzing the most important features:

The most important predictor for Small Forward players is pctFG, the percentage of field goals. It is s

The second most important predictor is pfTotals, the total number of personal fouls. It is significant

The third most important predictor is gp, the total number of games played. It is significant at 5% lev

Now, after exploring all the possible position, we created a table to store the significance informatio

	A	B	C	D	E	F	G	H	I
1	Group\Feature	Games played	Field Goal %	Free Throw %	Minutes played	Rebounds	Assists	Personal Fouls	Total points
2	Center		-5%						
3	Power Forward				-5%	1%			5%
4	Point Guard	-1%					1%	-10%	
5	Small Forward	5%	-1%					1%	
6	Shooting Guard	-5%	10%	1%					

First of all, the existence of entry represents that the given predictor is significant for the selection/suggestion of players from the given position. For example, the C2 entry stands for the result that field goal rate is significant for decisions among Center players.

Then, while the values stand for the significance level, the sign of entries stand for the direction of

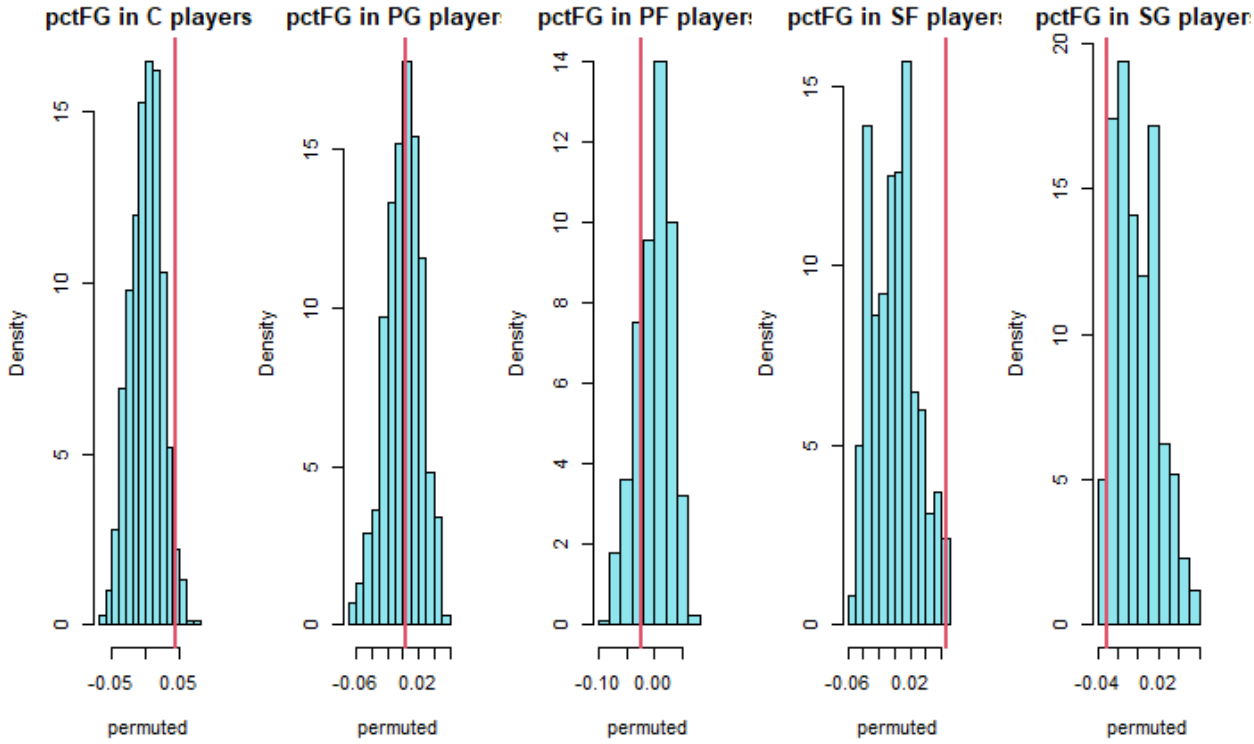
Now, based on these results from multiple linear regression, we could observe that the negative entries are overwhelming mainly on Games played and Field goal rate, while the positive entries seem to focus more on Rebounds and Assists.

(2) Permutation Test

Next, as we want to explore more on such relationship between statistics and the target variable, we decided to run hypothesis testing on the position/feature pairs, and the first test we worked on was the permutation test.

For the permutation test on a certain pair of position/feature, we start with the null hypothesis that the listed players and snubbed players, from given position, have a similar performance on the given feature. Then, as our target of hypothesis testing, our alternative is that the response variable Status is unevenly distributed along the given feature, so the actual difference in performance between the two groups is non-zero, and one group would have a generally higher level of performance for the position/feature pair.

Here's the R result example of our permutation tests, illustrating how we arrive at results.




```
[1] "The observed difference of pctFG is higher than 0.973 of the permuted
differences for C players"
[1] "The observed difference of pctFG is higher than 0.555 of the permuted
differences for PG players"
[1] "The observed difference of pctFG is higher than 0.213 of the permuted
differences for PF players"
[1] "The observed difference of pctFG is higher than 0.976 of the permuted
differences for SF players"
[1] "The observed difference of pctFG is higher than 0.042 of the permuted
differences for SG players"
```

The figures stand for the pairs of all positions and pctFG, the percentage of field goal rate. In each of the five plots, while the blue histogram represents the distribution of difference in mean pctFG for snubbed players

and listed players, the null hypothesis expects the distribution to be roughly normal centered at 0, while the vertical red lines stands for the observed difference in mean. (Actually, we subtract the mean of listed players from mean of snubbed players, so a left-skewed red line suggests that listed players generally have a higher mean, while a right-skewed red line supports the snubbed players.)

First of all, simply by looking at the figures, it is clear for us that the observed difference in mean

Then, we shall accompany the plots by the actual p-value expressed in the textual messages. And we coul

Now, the results from all permutation tests are also assembled into a table below:

	A	B	C	D	E	F	G	H	I
1	Group\Feature	Games played	Field Goal %	Free Throw %	Minutes played	Rebounds	Assists	Personal Fouls	Total points
2	All positions		-5%				5%		
3	Center		-5%						
4	Power Forward					10%	10%		
5	Point Guard								
6	Small Forward		-5%						
7	Shooting Guard		5%	10%					

It uses the same way to encode the entries, and we could conclude from the table that based on permutation test method, the players that are listed still have a better performance on team-work related statistics like Rebounds and Assists, while the players that are snubbed have a better performance on the field goal rate.

(3)T-test

We have also conducted the 2-sample T-test to test the difference in means of features on different positions. For a single pair we perform 2 steps of T-test to discover existence of difference in mean and validate its significance level. For Example:

```
> t.test(C_0$pctFG,C_1$pctFG,"two.sided")

        Welch Two Sample t-test

data:  C_0$pctFG and C_1$pctFG
t = 2.2008, df = 9.7798, p-value = 0.05296
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0006867615  0.0891855393
sample estimates:
mean of x mean of y
0.5402193 0.4959699
```

When testing for field goal rate and Center pair, our two-sided test suggests a p-value of around 0.05. As we found that most of the two-sided T-tests were not able to arrive at a significance level of 5%, we decided to loosen our criterion a bit for T-test models and used a deciding threshold of 10% significance level for two-sided tests. Thus, from this figure we could see that the difference in mean of field goal rate is significant on 10% level, and our observed direction of such difference is positive(0.540>0.495), indicating that snubbed Center players are likely to have a higher average field goal rate than listed Center players.

Then, in order to investigate the exact significance level of such difference, we ran a second-stage T-

```
> t.test(C_0$pctFG,C_1$pctFG,"greater")

        Welch Two Sample t-test

data:  C_0$pctFG and C_1$pctFG
t = 2.2008, df = 9.7798, p-value = 0.02648
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.007724536         Inf
sample estimates:
mean of x mean of y
0.5402193 0.4959699
```

The p-value is now 0.02648, and the 95% confidence interval does not include 0 which stands for the null hypothesis, so we could conclude from this test that snubbed Center players are likely to have a higher level of field goal rate than the listed Center players, somehow supporting the hypothesis that field goal rate is suggested to be underestimated as a factor of the 75th Anniversary List selection.

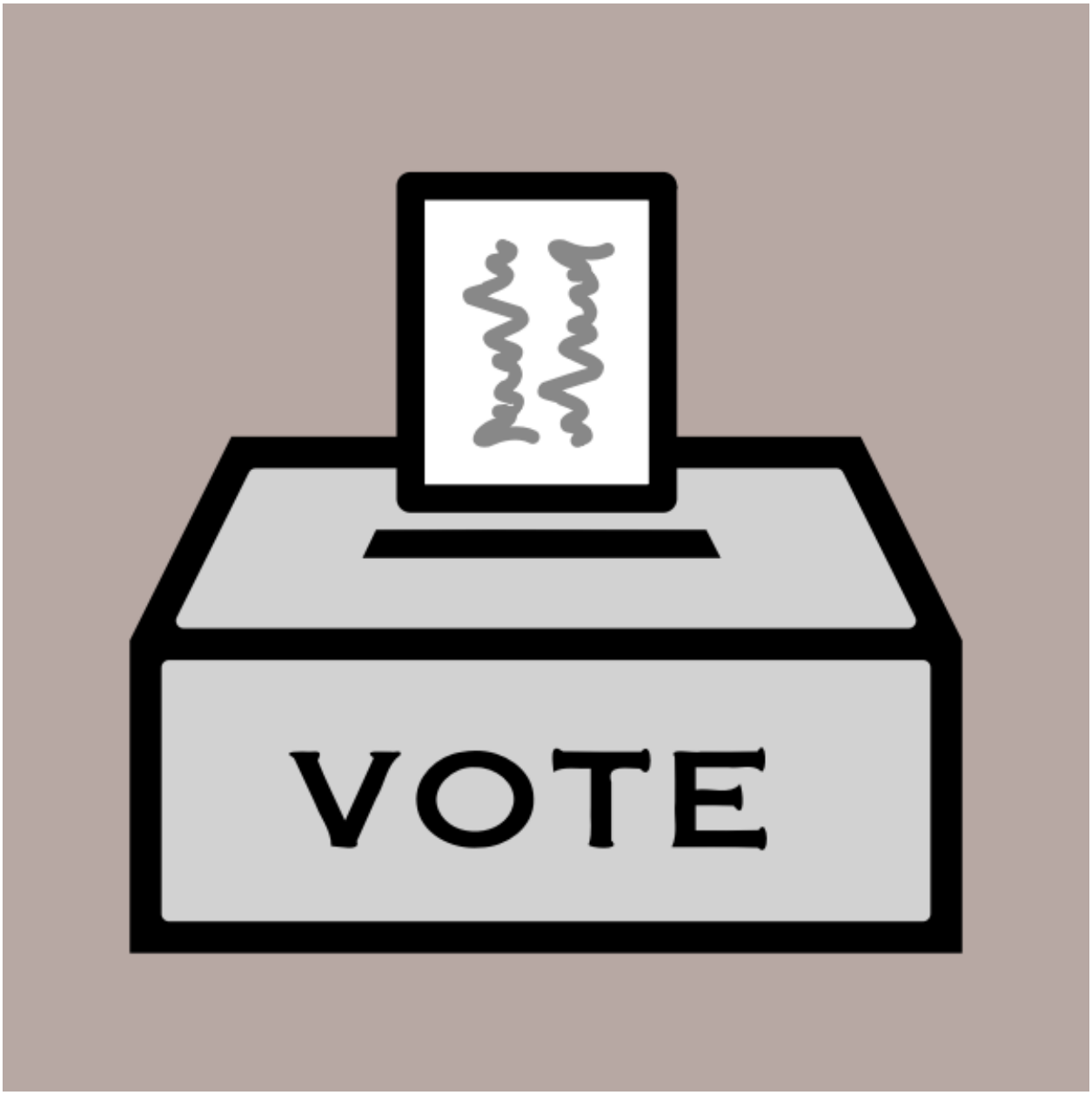
Now, we also have the result chart of these T-tests below:

	A	B	C	D	E	F	G	H	I
1	Group\Feature	Games played	Field Goal %	Free Throw %	Minutes played	Rebounds	Assists	Personal Fouls	Total points
2	All positions						1%		5%
3	Center		-5%						
4	Power Forward					5%	5%	-5%	
5	Point Guard						5%	-5%	
6	Small Forward		-5%				5%		
7	Shooting Guard		1%						

We could observe that while snubbed players seem to be better at field goals and personal fouls(making less fouls), the listed players mainly have their advantages at teamworking statistics like rebounds and assists.

3. Generalization of the Results

Throughout the project, we have utilized 3 different methods on the position/feature pairs of our dataset. While the general patterns of the 3 result charts seems similar, their exact results are different from each other. Thus, in order to arrive at conclusions solid enough to be supported by our methods. We decided to let the 3 methods "vote" for results. This means that for a certain position/feature pair, if at least 2 of the 3 methods suggest that listed players have a better performance on the pair, then we agree on the result and conclude that the given feature could be important for selection of 75th Anniversary List players from the given position. However, if at least 2 of the 3 methods suggest that snubbed players have a better performance, then we would also agree and conclude that the given feature is suggested to be underestimated for the selection of listed players from the given position.



Based on this voting method, we aggregated the 3 charts and arrived at the generalized chart of importance below:

	A	B	C	D	E	F	G	H	I
1	Group\Feature	Games played	Field Goal %	Free Throw %	Minutes played	Rebounds	Assists	Personal Fouls	Total points
2	All positions						Important		
3	Center		Underestimated						
4	Power Forward					Important	Important		
5	Point Guard						Important	Underestimated	
6	Small Forward		Underestimated				Important		
7	Shooting Guard		Important	Important					

The number of important position/factor pairs(7) is much more than the number of underestimated pairs(3). This seems to suggest that the overall performance of listed players were further accepted by the 3 methods, but more important inferences still need to be made.

Conclusions

Now, let us conclude a bit on the suggestion effectiveness of the snubbed players. The field goal rate and personal foul are the most underestimated features prompted by the suggestions, and these 2 factors usually relate closely to the offensive behaviors of a NBA player, so it seems reasonable to infer that the suggestions of those snubbed players mainly focus their offensiveness of players while looking down on the rest of their skills.

Then, it is time to derive some useful conclusions on the results we have reached. Why would suggestion focus mainly on offensiveness while the 75th Anniversary List considers more on the other skills such as rebounds and assists?

First of all, we want to hypothesize that the aggressiveness would be important for the direct impression a NBA player would left for their fans: An accurate field goal usually leads to a great impression on the player's personal performance, and the ability to make less fouls could further improve the perfectness of such impression. In this logic, as a player keeps making field goals without personal fouls, they would be likely to leave a stronger impression for the audience, and to be mentioned as a snubbed player when the audience compare the actual list and find their NBA hero not appearing on the list.

Then, for listed players, their advantages comparing to the snubbed players seem to focus mainly on rebounds and assists, which demonstrate another important factor for the success of the whole team: teamwork. While the goals are generally more eyeball-catching, the teamwork behind those goals serves as the solid basis and takes an important role in the NBA games. Just imagine yourself as a NBA player mainly focusing on making goals. How often would you finish a goal solely carried out by yourself? Scrambling on the control of ball, passing the ball to teammates at better position, preventing opponents from making a goal... There are really many actions and skills that determine the overall performance of a NBA team, and a victory usually comes mostly from the solid support your teammates provide to you. Based on these inferences, while the goal-focused suggestions of snubbed players might come from the direct impressions of normal audiences, it seems natural for the 75th Anniversary List to consider more in the overall capabilities of a NBA player, like the rebound and assist skills.

Thus, for the final conclusions we have drawn from such exploration of possible factors:

(1) Contrasting to personal performance factors, it is likely that the 75th Anniversary List players were selected for their overall skills to cooperate with teammates and their importances for their whole team, like rebounds and assists. However, if these teamwork factors were not the most impressive factors for normal audience, then for list makers there might exist the risk that our Anniversary List ultimately becomes something too professional for the common audience to understand, to agree with, and to give their advice on the improvement of the list.

(2) The suggestion of snubbed players seems to focus mainly on offensiveness factors like field goal rate and personal fouls, which would seem natural if these factors were indeed the strongest factors that could leave an impression of the audience. Thus, for the NBA game audience, it would be a good idea to focus on more integral skill sets for the NBA players, which in turn makes people more "professional" when commenting on players.

(3) For commentators delivering a commentary to the NBA games, because audience possibly favors players with better offensive skills, if they could shift their commenting styles to focus more on personal performance and highlight players making the goals, then the audience is likely to get affected and become more passionate.

However, there are also some limitations of our project. First of all, since many players have multiple main positions and each of them has to be in only one subset as the total number of observations is rather small, there could be some issues on the conditional distributions. Secondly, the dataset itself is not so balanced on the response variable Status, which might also impact the overall reliability of our results. Also, the NBA rules could be different based on different periods, so merely looking at those statistics could possibly lead

us to overlook time as another important but ignored factor of the list selection. Plus, as our conclusions were mainly based on the hypothesis of NBA audience focusing on personal performance, these conclusions are somewhat unstable and could be challenged greatly if the hypothesis itself is challenged to be insolid.

Overall, the project has utilized lots of reliable methods, and results have expressed a general idea on the importance of different factors in the establishment of the 75th anniversary list. While the focus of list selection and suggestion on snubbed players seem to have different preference over the factors, a critical hypothesis of audience focusing on personal performance might be applied to explain these phenomenon and help derive meaningful conclusions for the list makers, for commentators, and for our audience.

Now, it is time to introduce some of our future research directions. While the hypothesis that common audience would be likely to focus on personal performance needs further exploration and validation, we also have another possible direction to reach our target: while we currently have focused on the statistical comparisons between snubbed players and listed players, one of our alternative strategies would be to develop an alternative list of 75 players, which requires us to replace some of the currently listed players with snubbed players. Then, in order to explore possibilities of such replacements, one available method could be the Bayesian model that if player B SHOULD replace player A in the list, then the probability of model predicting B in list given A is in list would be larger than the probability of model predicting A in list given B replaces A in the list. By this method, given a snubbed player and a player in the list, we could use Bayesian models such as Naive Bayes to compute probabilities and see whether the snubbed one is truly worthy to be in the list.

However, 4 concerns still need to be addressed here.

First of all, for a certain snubbed player, who exactly should be considered as their "chivalry" to replace? Although NBA reporters and analysts certainly have their opinions, those are hard to collect, assemble and analyze. Since our dataset mainly consists of player statistics, we shall assume these are sufficient criterions for presence in the list; then, for a certain player outside the list, although there could be many other choices, we could find the player in the list most similar to them, that is, having the smallest distance from them in the statistics.

Secondly, such a method only models replacement one by one, so should we consider the previous replacements traversing the next player to replace? Here, our answer is "No", since there's no such a determined sequence of snubbed players to traverse, and different routes to traverse could lead to different results if we really update the list for each replacement; also, the opinions for snubbed players are based on the current stage of the list. Thus, for robustness of the replacements, we shall fix the list upon all replacements.

Thirdly, if we really decide to make the comparison for all of the snubbed players one by one, what if there are 2 snubbed players having similar statistics that they are to replace the same one in the list? Again there could be many ways to approach this, but our current expectation would be to run a second comparison between the "qualified" snubbed players and decide who would contribute to the final replacement.

Finally, how should we interpret such replacements? Could there be a chance that a snubbed player makes it to replace a player in the list who has literally better statistics than them? This would also be a target to explore in the future.

Appendix

```
player_data <- read.csv("AllPlayerStats.csv",row.names = 1)
```

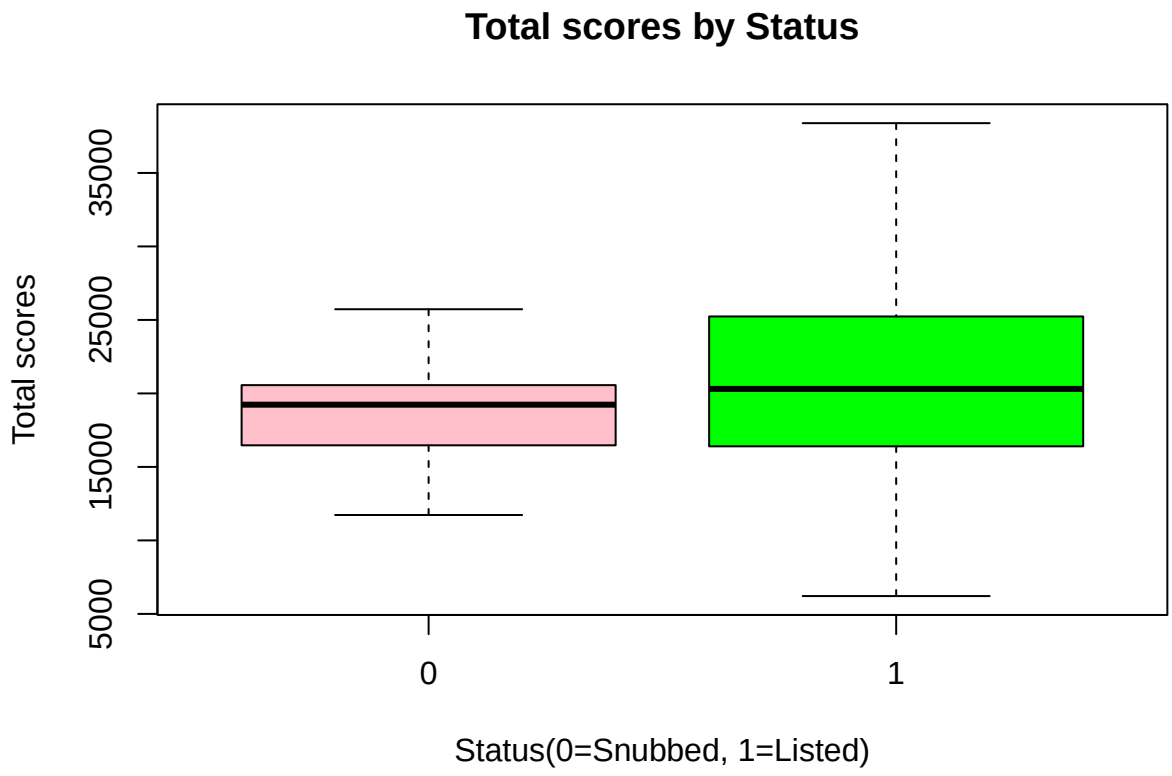
Dividing into subsets of different positions

```
SF <- player_data[player_data$Position == "SF",][-c(11)]
PG <- player_data[player_data$Position == "PG",][-c(11)]
PF <- player_data[player_data$Position == "PF",][-c(11)]
C <- player_data[player_data$Position == "C",][-c(11)]
SG <- player_data[player_data$Position == "SG",][-c(11)]
```

Exploratory Data Analysis

1. Boxplots of total scores (ptsTotals) for listed players and snubbed players

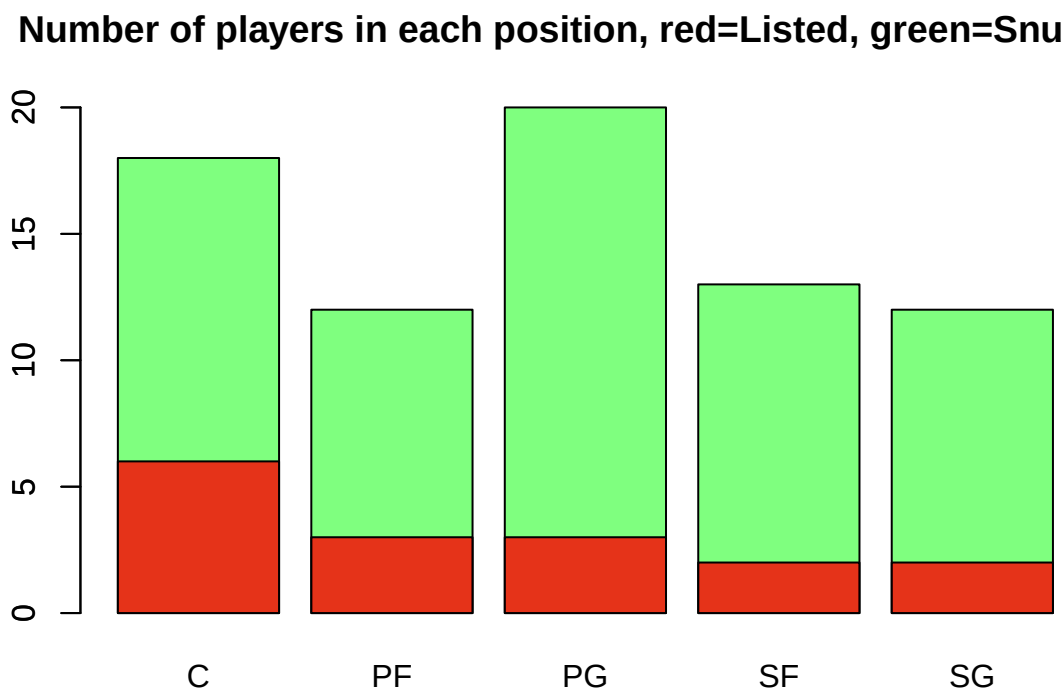
```
boxplot(ptsTotals~Status,data=player_data, main="Total scores by Status",
   xlab="Status(0=Snubbed, 1=Listed)", ylab="Total scores",col=c('pink','green'))
```



No obvious difference in median of total points.

2. The number of players in each position

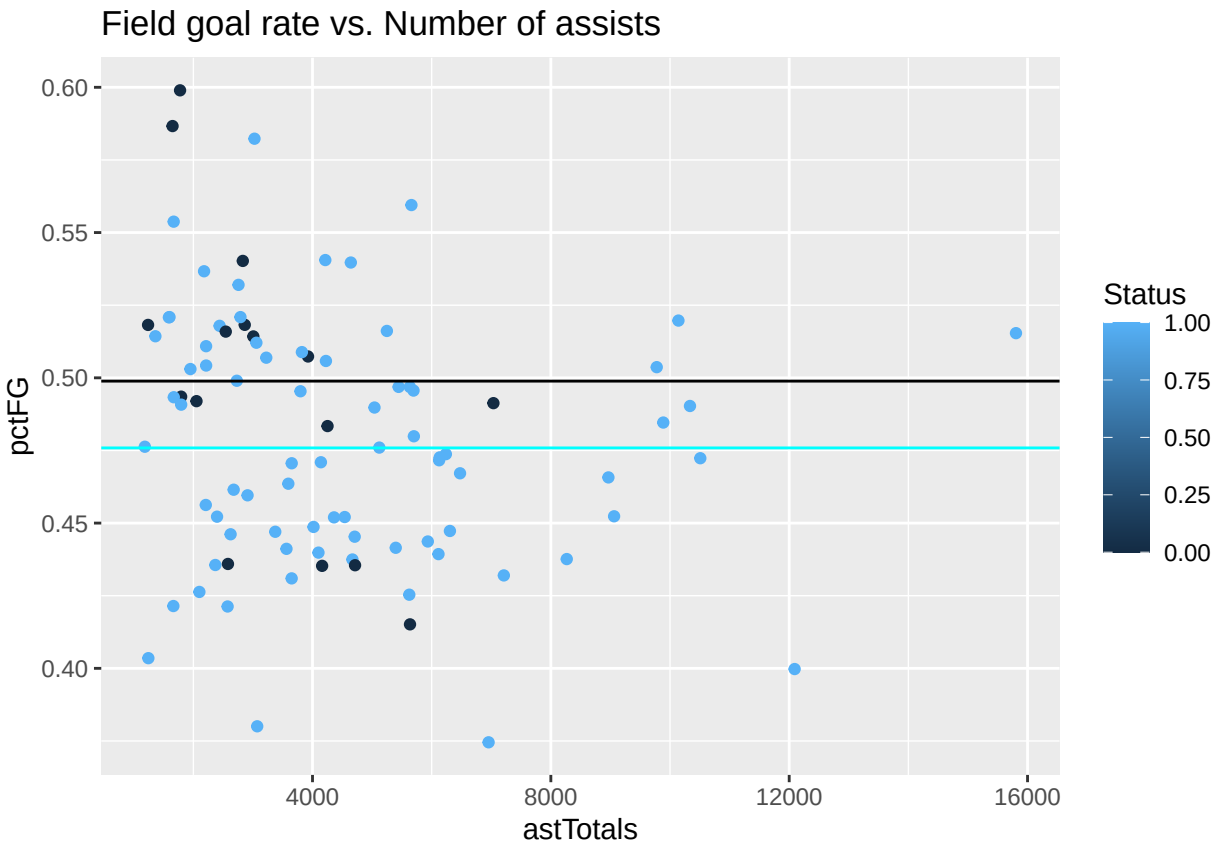
```
snubbed <- player_data[player_data$Status == 0,]
listed <- player_data[player_data$Status == 1,]
listed_count <- aggregate(Name ~ Position, data=listed, function(x) length(unique(x)))$Name
snubbed_count <- aggregate(Name ~ Position, data=snubbed, function(x) length(unique(x)))$Name
barplot(listed_count, col=rgb(0, 1, 0, .5),main="Number of players in each position, red=Listed, green=
barplot(snubbed_count, col=rgb(1, 0, 0, 0.8), add=TRUE)
```



From this plot, we could observe that the both the absolute and relative size of snubbed players are small, roughly 1/6~1/3 of that for listed players in each of the 5 positions, so the analysis results might be a bit unstable.

3. Scatter plot of field goal rate against number of assists

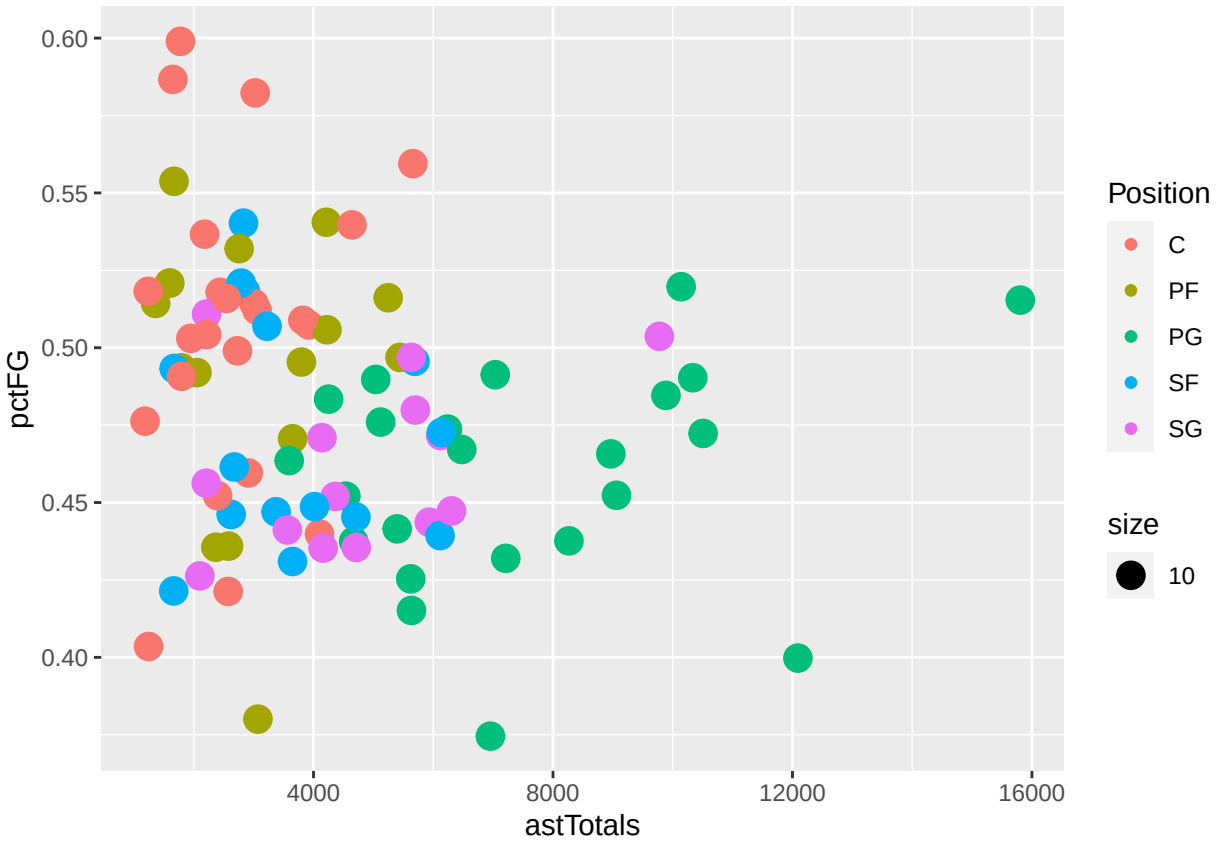
```
library(ggplot2)
qplot(x = astTotals, y = pctFG, data = player_data, color=Status) + geom_hline(yintercept=mean(snubbed$
```



From this plot we can see that while listed players seem to have a wider distribution of number of assists and many players well at assistance, the observed snubbed players seem to have a higher average in field goal rates, but such relationships would need further validation to be properly argued.

4. relative advantages of each position

```
qplot(x = astTotals, y = pctFG, data = player_data, color=Position) + geom_point(aes(colour = Position,
```



From plot we could see that while Center players generally have the highest level of field goal accuracy, Point Guard players seem to focus on teamwork instead of being a scorer, while the rest 3 positions could not be separated well from the current stage of data we obtained.

Exploring Short Forward players

Method 1: Multiple linear regression

```
fit_SF <- lm(Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals + astTotals + pfTotals + ptsTotals
summary(fit_SF)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals +
##     astTotals + pfTotals + ptsTotals, data = SF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23215 -0.06935 -0.01154  0.05298  0.36309
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.720e+00  2.638e+00   1.031   0.3421
## gp             5.575e-03  3.428e-03   1.626   0.1550
## pctFG         -6.800e+00  2.544e+00  -2.673   0.0369 *
```

```
## pctFT          1.408e+00  1.832e+00   0.769   0.4713
## minutesTotals -1.418e-04  1.335e-04  -1.062   0.3291
## trebTotals     1.363e-04  8.614e-05   1.582   0.1647
## astTotals      8.688e-05  1.388e-04   0.626   0.5544
## pfTotals      -3.822e-04  1.822e-04  -2.097   0.0808 .
## ptsTotals     -2.689e-05  6.436e-05  -0.418   0.6907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.225 on 6 degrees of freedom
## Multiple R-squared:  0.8247, Adjusted R-squared:  0.591
## F-statistic: 3.529 on 8 and 6 DF,  p-value: 0.07089
```

```
fit_SF_back <- lm(Status ~ gp + pctFG + minutesTotals + trebTotals + astTotals + pfTotals,data=SF)
summary(fit_SF_back)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + minutesTotals + trebTotals +
##     astTotals + pfTotals, data = SF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29324 -0.07662 -0.04086  0.10540  0.32926
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.533e+00  1.072e+00   4.228  0.00288 **
## gp             5.298e-03  2.020e-03   2.623  0.03052 *
## pctFG         -7.933e+00  1.850e+00  -4.287  0.00266 **
## minutesTotals -1.472e-04  6.095e-05  -2.415  0.04218 *
## trebTotals     1.008e-04  6.239e-05   1.616  0.14486
## astTotals      1.367e-04  5.863e-05   2.332  0.04800 *
## pfTotals      -4.614e-04  1.362e-04  -3.388  0.00952 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2044 on 8 degrees of freedom
## Multiple R-squared:  0.8072, Adjusted R-squared:  0.6627
## F-statistic: 5.584 on 6 and 8 DF,  p-value: 0.01483
```

The most important features for Short Forward players are pctFG,pfTotals and gp.

Listed SF players generally have lower pctFG, lower pfTotals and higher gp, lower minutesTotals, higher astTotals

```
fit_SG <- lm(Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals + astTotals + pfTotals + ptsTotals
summary(fit_SG)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals +
##     astTotals + pfTotals + ptsTotals, data = SG)
```

```
##
## Residuals:
##          7         19         30         35         42         56         59
##  0.0197886  0.1054265  0.1709196 -0.1703538 -0.2117437  0.3093755  0.0019710
##         62         73         74         75         79         86         87
## -0.1208346  0.0843296 -0.0006737 -0.0374502  0.3752357 -0.3449938 -0.1809968
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.696e+00  3.795e+00  -1.765    0.138
## gp            -4.694e-04  1.976e-03  -0.238    0.822
## pctFG          7.202e+00  4.551e+00   1.583    0.174
## pctFT          5.641e+00  3.170e+00   1.780    0.135
## minutesTotals  8.920e-07  6.487e-05   0.014    0.990
## trebTotals    -1.429e-05  1.563e-04  -0.091    0.931
## astTotals      1.449e-04  1.309e-04   1.107    0.319
## pfTotals      -4.861e-05  2.394e-04  -0.203    0.847
## ptsTotals     -1.834e-05  4.778e-05  -0.384    0.717
##
## Residual standard error: 0.3243 on 5 degrees of freedom
## Multiple R-squared:  0.6932, Adjusted R-squared:  0.2022
## F-statistic: 1.412 on 8 and 5 DF,  p-value: 0.3661
```

```
fit_SG_back <- lm(Status ~ gp + pctFG + pctFT + astTotals,data=SG)
summary(fit_SG_back)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + pctFT + astTotals, data = SG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34733 -0.20226  0.03781  0.12138  0.38387
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.7858697  1.8714490  -3.092  0.01290 *
## gp          -0.0007559  0.0003122  -2.421  0.03853 *
## pctFG        5.7999413  2.7352244   2.120  0.06299 .
## pctFT        5.2136622  1.5953810   3.268  0.00971 **
## astTotals    0.0001026  0.0000492   2.085  0.06670 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2506 on 9 degrees of freedom
## Multiple R-squared:  0.6702, Adjusted R-squared:  0.5236
## F-statistic: 4.572 on 4 and 9 DF,  p-value: 0.02729
```

Listed SG players tend to have greater pctFT, less gp.

```
fit_C <- lm(Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals + astTotals + pfTotals + ptsTotals
summary(fit_C)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals +
##     astTotals + pfTotals + ptsTotals, data = C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76035 -0.27591  0.09687  0.24300  0.48671
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.592e+00  2.463e+00   1.053    0.309
## gp            -1.949e-04  1.225e-03  -0.159    0.876
## pctFG         -4.193e+00  2.658e+00  -1.577    0.136
## pctFT          2.401e-02  1.880e+00   0.013    0.990
## minutesTotals -2.844e-05  5.246e-05  -0.542    0.596
## trebTotals     3.333e-05  7.138e-05   0.467    0.647
## astTotals      4.747e-05  1.302e-04   0.365    0.721
## pfTotals       8.951e-05  2.427e-04   0.369    0.717
## ptsTotals      3.414e-05  2.355e-05   1.450    0.168
##
## Residual standard error: 0.4518 on 15 degrees of freedom
## Multiple R-squared:  0.3196, Adjusted R-squared:  -0.04332
## F-statistic: 0.8806 on 8 and 15 DF,  p-value: 0.554
```

```
fit_C_back <- lm(Status ~ pctFG + ptsTotals,data=C)
summary(fit_C_back)
```

```
##
## Call:
## lm(formula = Status ~ pctFG + ptsTotals, data = C)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7718 -0.1734  0.1385  0.2549  0.5625
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.788e+00  8.797e-01   3.169  0.00462 **
## pctFG       -4.733e+00  1.845e+00  -2.565  0.01805 *
## ptsTotals    1.843e-05  1.157e-05   1.593  0.12603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4006 on 21 degrees of freedom
## Multiple R-squared:  0.2511, Adjusted R-squared:  0.1797
## F-statistic:  3.52 on 2 and 21 DF,  p-value: 0.04805
```

Listed Center players tend to have lower pctFG.

```
fit_PF <- lm(Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals + astTotals + pfTotals + ptsTotals
summary(fit_PF)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals +
##     astTotals + pfTotals + ptsTotals, data = PF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45868 -0.12643 -0.00814  0.09850  0.38017
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.518e+00  4.154e+00  -0.847   0.4295
## gp             3.886e-03  2.301e-03   1.689   0.1422
## pctFG          5.944e+00  3.608e+00   1.648   0.1505
## pctFT          1.105e+00  3.369e+00   0.328   0.7541
## minutesTotals -1.904e-04  7.860e-05  -2.422   0.0517 .
## trebTotals     1.859e-04  8.471e-05   2.195   0.0706 .
## astTotals      1.106e-04  1.289e-04   0.858   0.4238
## pfTotals      -1.813e-04  3.023e-04  -0.600   0.5706
## ptsTotals      7.441e-05  5.041e-05   1.476   0.1904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3508 on 6 degrees of freedom
## Multiple R-squared:  0.6923, Adjusted R-squared:  0.2819
## F-statistic: 1.687 on 8 and 6 DF,  p-value: 0.2702
```

```
fit_PF_back <- lm(Status ~ gp + pctFG + minutesTotals + trebTotals + ptsTotals,data=PF)
summary(fit_PF_back)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + minutesTotals + trebTotals +
##     ptsTotals, data = PF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54182 -0.11612 -0.05686  0.17610  0.45390
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.227e+00  1.103e+00  -2.019  0.07420 .
## gp             3.341e-03  1.622e-03   2.060  0.06945 .
## pctFG          4.990e+00  2.100e+00   2.376  0.04147 *
## minutesTotals -1.736e-04  6.356e-05  -2.731  0.02319 *
## trebTotals     1.462e-04  4.353e-05   3.358  0.00842 **
## ptsTotals      8.607e-05  3.239e-05   2.657  0.02616 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3172 on 9 degrees of freedom
## Multiple R-squared:  0.6226, Adjusted R-squared:  0.4129
## F-statistic: 2.969 on 5 and 9 DF,  p-value: 0.07431
```

Listed Power Forward players tend to have higher rebounds, lower minutesTotals, higher ptsTotals and pctFG.

```
fit_PG <- lm(Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals + astTotals + pfTotals + ptsTotals
summary(fit_PG)
```

```
##
## Call:
## lm(formula = Status ~ gp + pctFG + pctFT + minutesTotals + trebTotals +
##     astTotals + pfTotals + ptsTotals, data = PG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64972 -0.05816  0.03016  0.14596  0.38837
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.602e+00  1.714e+00   0.935   0.3658
## gp            -2.529e-03  1.025e-03  -2.468   0.0271 *
## pctFG         -1.543e+00  1.971e+00  -0.783   0.4466
## pctFT          2.673e-01  1.574e+00   0.170   0.8676
## minutesTotals  3.794e-05  3.810e-05   0.996   0.3362
## trebTotals    -4.058e-05  4.945e-05  -0.821   0.4256
## astTotals      8.383e-05  3.723e-05   2.252   0.0409 *
## pfTotals       2.263e-04  1.408e-04   1.607   0.1304
## ptsTotals     -4.398e-07  3.072e-05  -0.014   0.9888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3055 on 14 degrees of freedom
## Multiple R-squared:  0.499,  Adjusted R-squared:  0.2127
## F-statistic: 1.743 on 8 and 14 DF,  p-value: 0.1735
```

```
fit_PG_back <- lm(Status ~ gp + astTotals + pfTotals,data=PG)
summary(fit_PG_back)
```

```
##
## Call:
## lm(formula = Status ~ gp + astTotals + pfTotals, data = PG)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6084 -0.1616  0.0630  0.1508  0.3293
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.317e+00  3.394e-01   3.879  0.00101 **
## gp          -1.665e-03  4.985e-04  -3.339  0.00345 **
## astTotals    8.521e-05  2.901e-05   2.937  0.00846 **
## pfTotals     2.525e-04  1.207e-04   2.092  0.05011 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

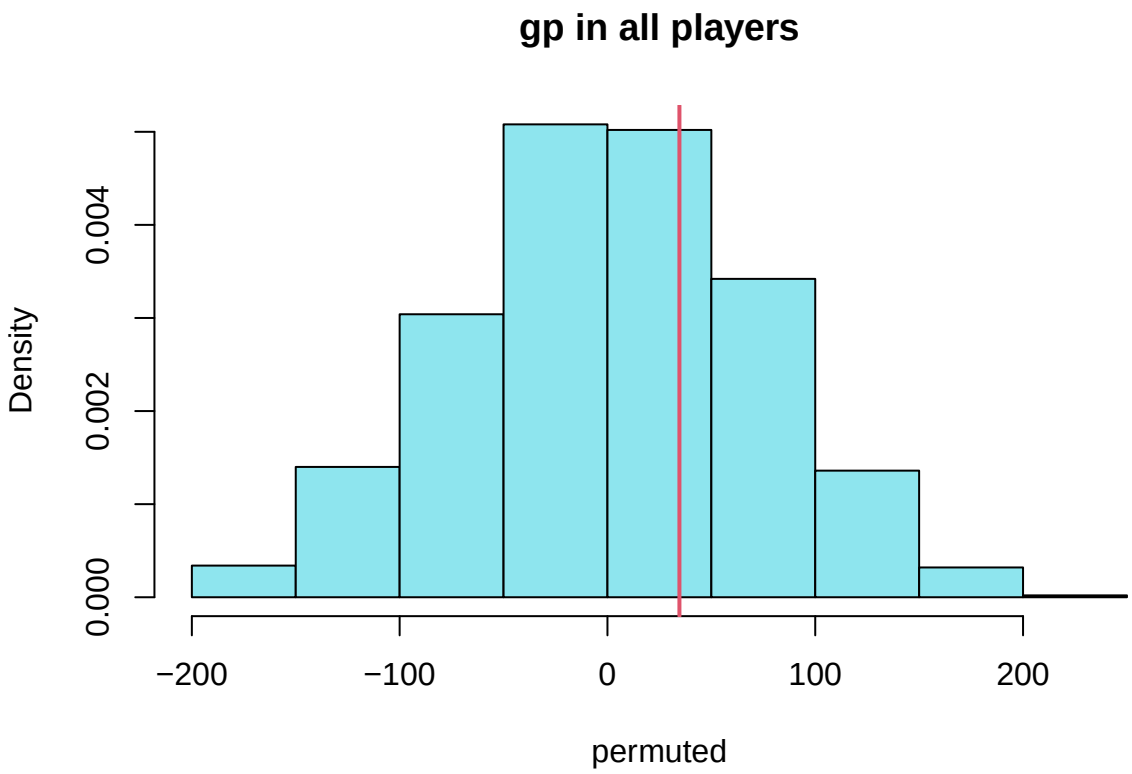
```
## Residual standard error: 0.2836 on 19 degrees of freedom
## Multiple R-squared:  0.4142, Adjusted R-squared:  0.3217
## F-statistic: 4.477 on 3 and 19 DF,  p-value: 0.01539
```

Listed Point Guard players tend to have lower gp, higher astTotals and higher pfTotals.

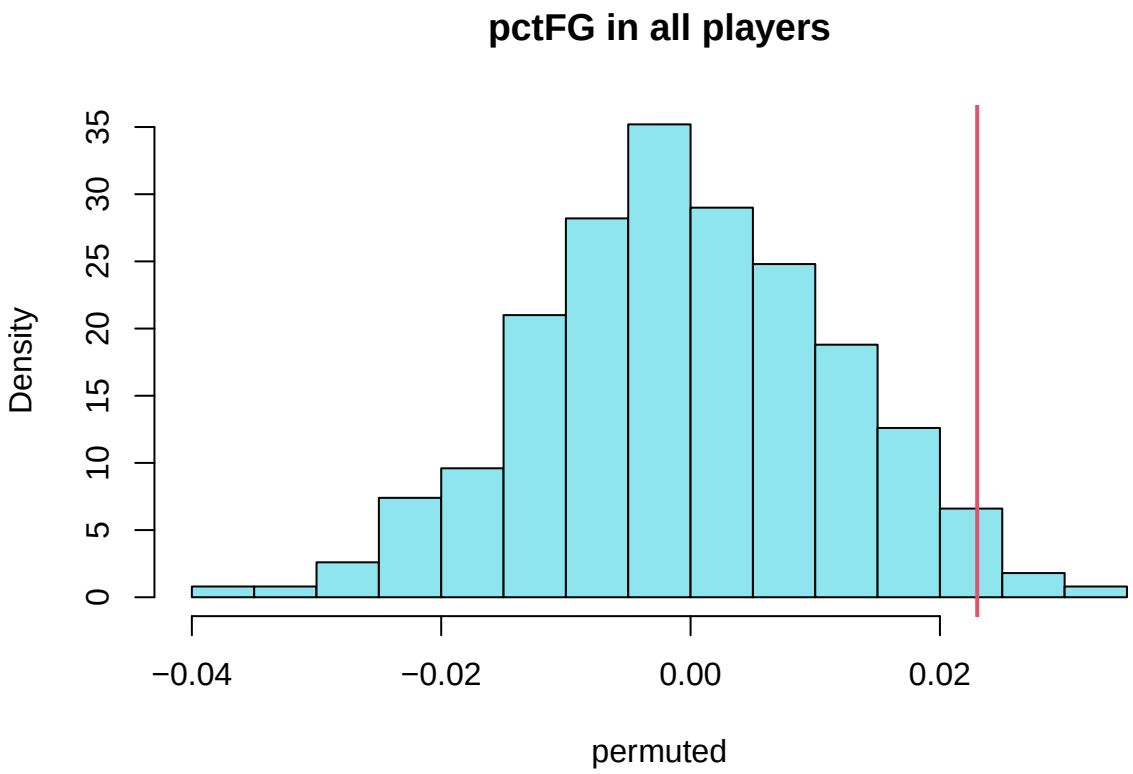
Method 2: Permutation tests

```
mean_diff <- function(df,feature) {
  agg = aggregate(formula(paste0(feature,"~Status")), data=df, FUN=mean)
  return(agg[,which(colnames(agg) == feature)][1] - agg[,which(colnames(agg) == feature)][2])
}
permute <- function(data,feature){
  permutation <- data.frame(data)
  permutation$Status = permutation$Status[sample(nrow(data),nrow(data),replace=F)]
  results <- c()
  results <- append(results,mean_diff(permutation,feature))
  return(results)
}
```

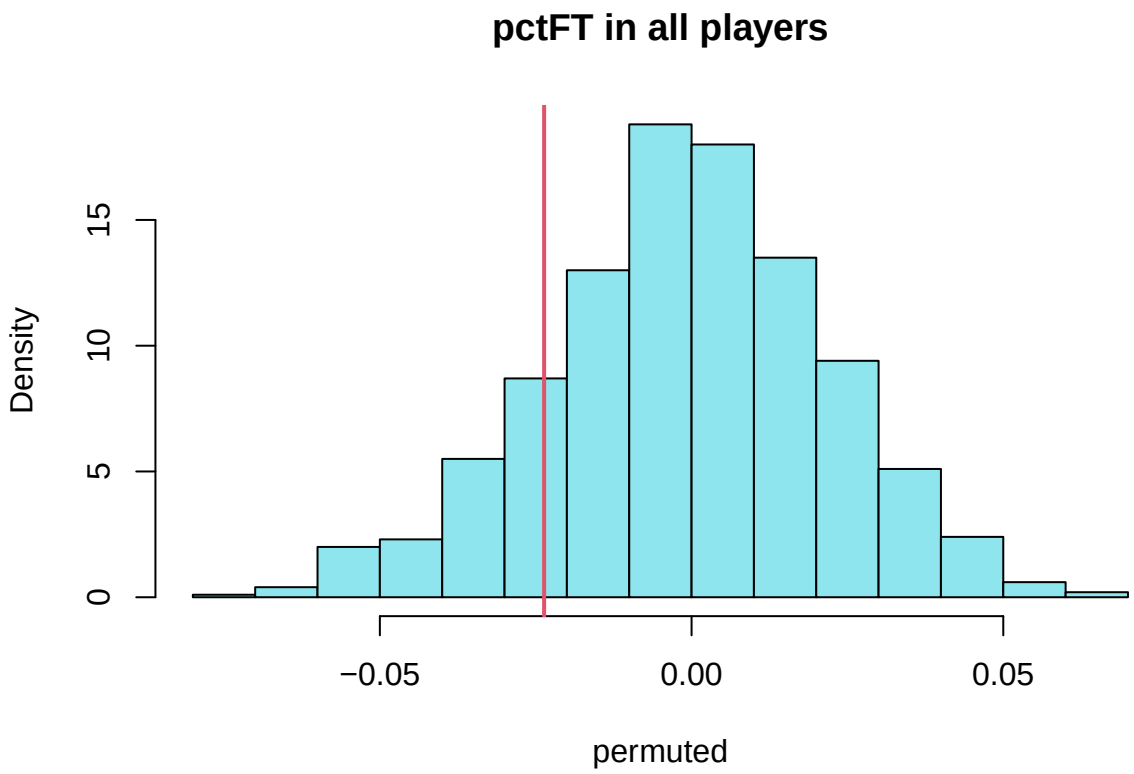
```
set.seed(10)
for(j in c("gp","pctFG","pctFT","minutesTotals","trebTotals","astTotals","pfTotals","ptsTotals")) {
    permuted <- replicate(1000,permute(player_data,j))
    observed <- mean_diff(player_data,j)
    hist(permuted, main = paste(j,"in","all","players",sep=" ") , prob=T, col="cadetblue2")
    abline(v = observed , col = 2, lwd = 2)
    print(paste0("The observed difference is higher than ",mean(permuted<observed)," of the permuted di
}
```



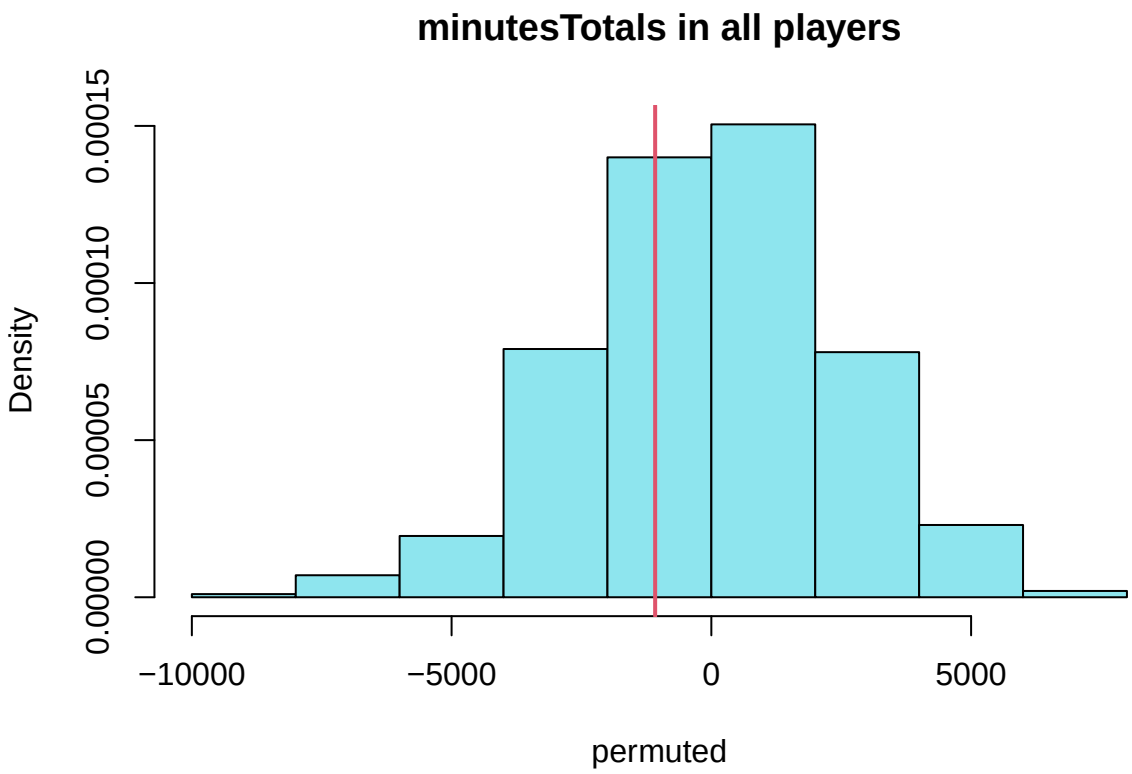

```
## [1] "The observed difference is higher than 0.679 of the permuted differences in gp"
```



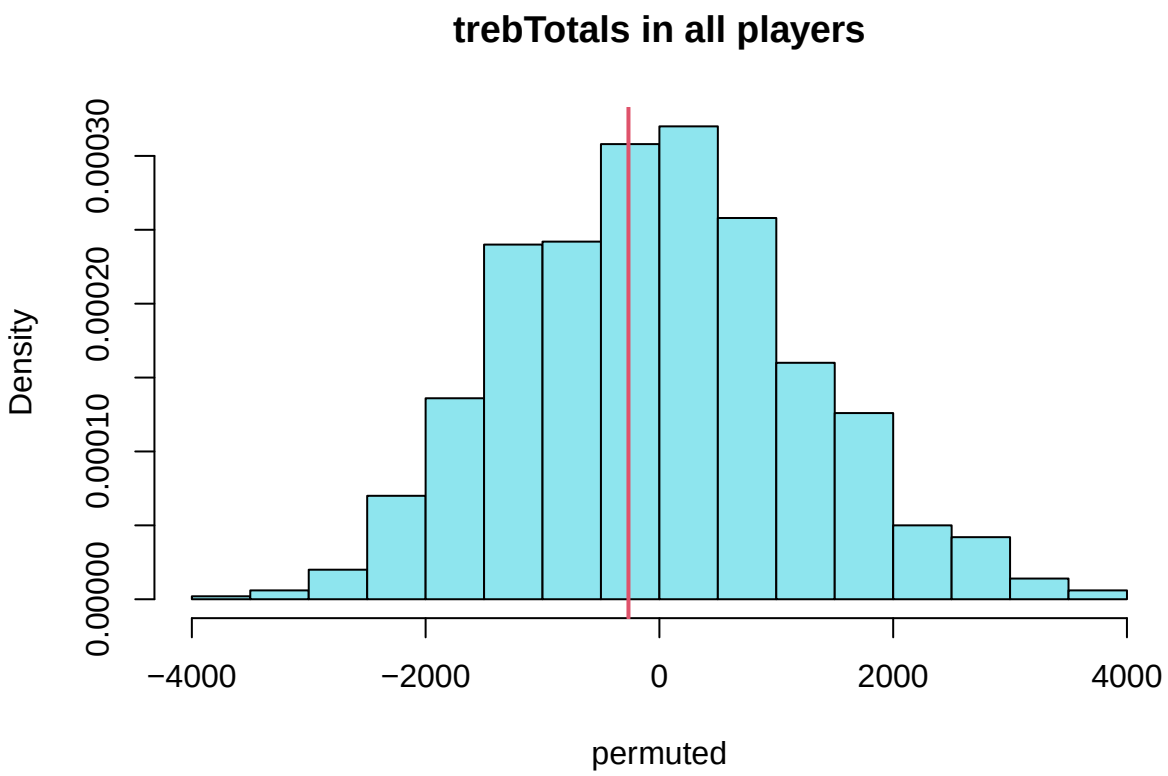

```
## [1] "The observed difference is higher than 0.979 of the permuted differences in pctFG"
```



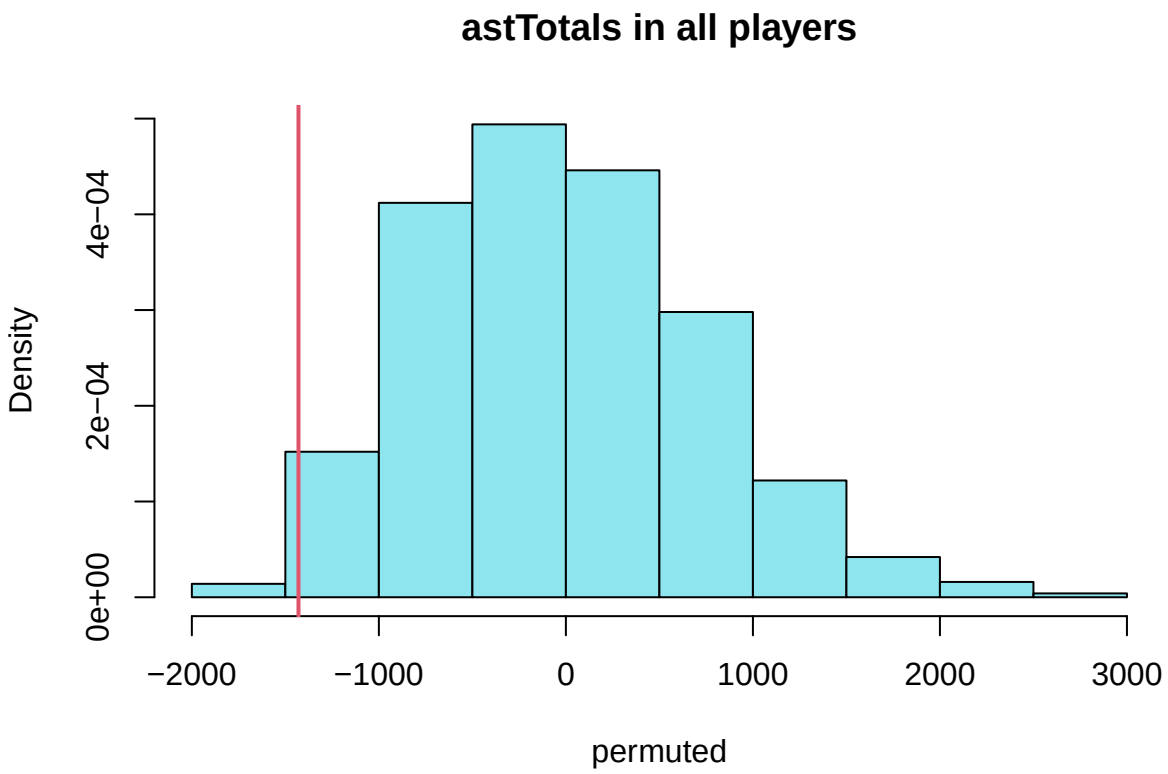

```
## [1] "The observed difference is higher than 0.157 of the permuted differences in pctFT"
```



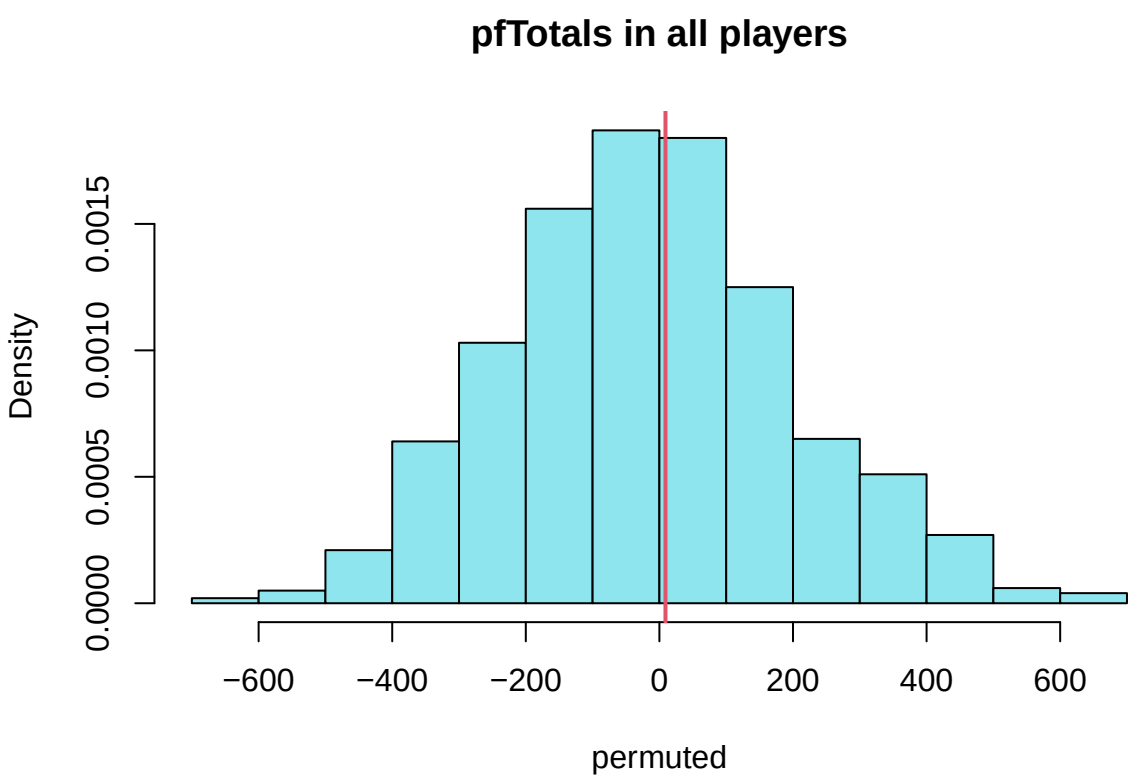

```
## [1] "The observed difference is higher than 0.346 of the permuted differences in minutesTotals"
```



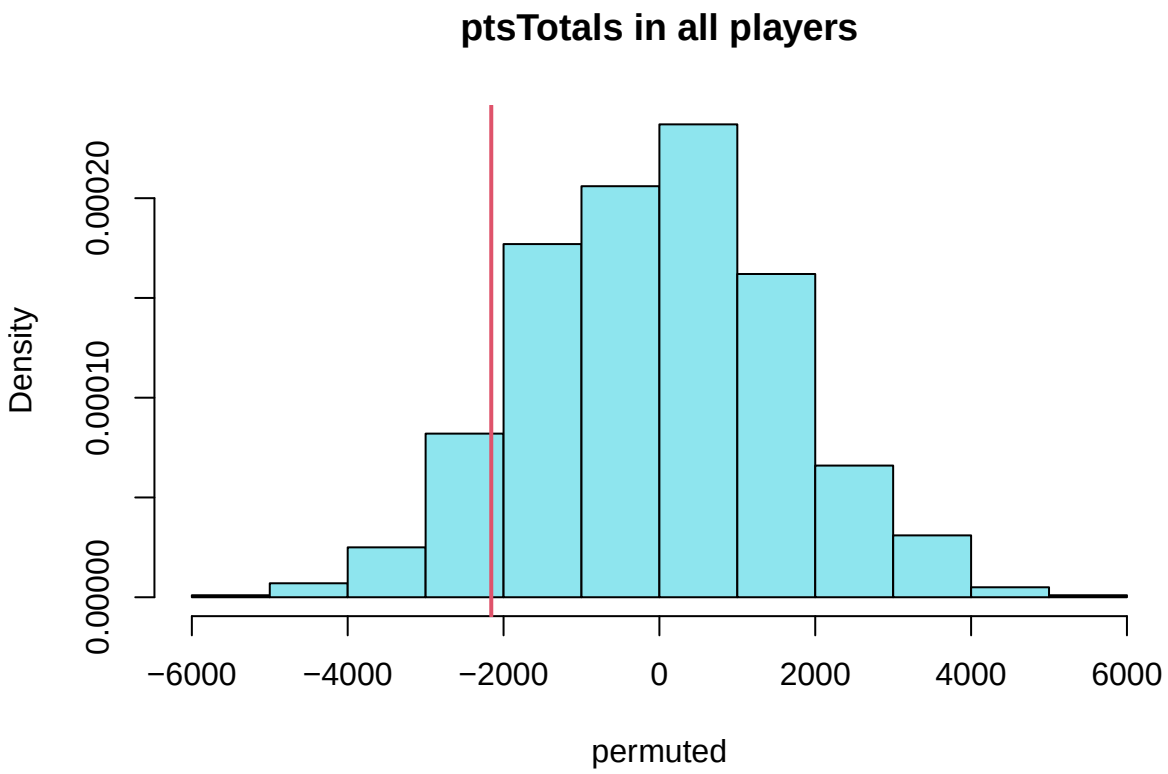

```
## [1] "The observed difference is higher than 0.439 of the permuted differences in trebTotals"
```




```
## [1] "The observed difference is higher than 0.016 of the permuted differences in astTotals"
```




```
## [1] "The observed difference is higher than 0.553 of the permuted differences in pfTotals"
```




```
## [1] "The observed difference is higher than 0.103 of the permuted differences in ptsTotals"
```

Considering players from all positions, the snubbed players have a significantly higher average pctFG and significantly lower average astTotals, on the 5% significance level. Also, although not as significant as the above 2, ptsTotals is another factor worths attention, as it is almost significant at 10% level.

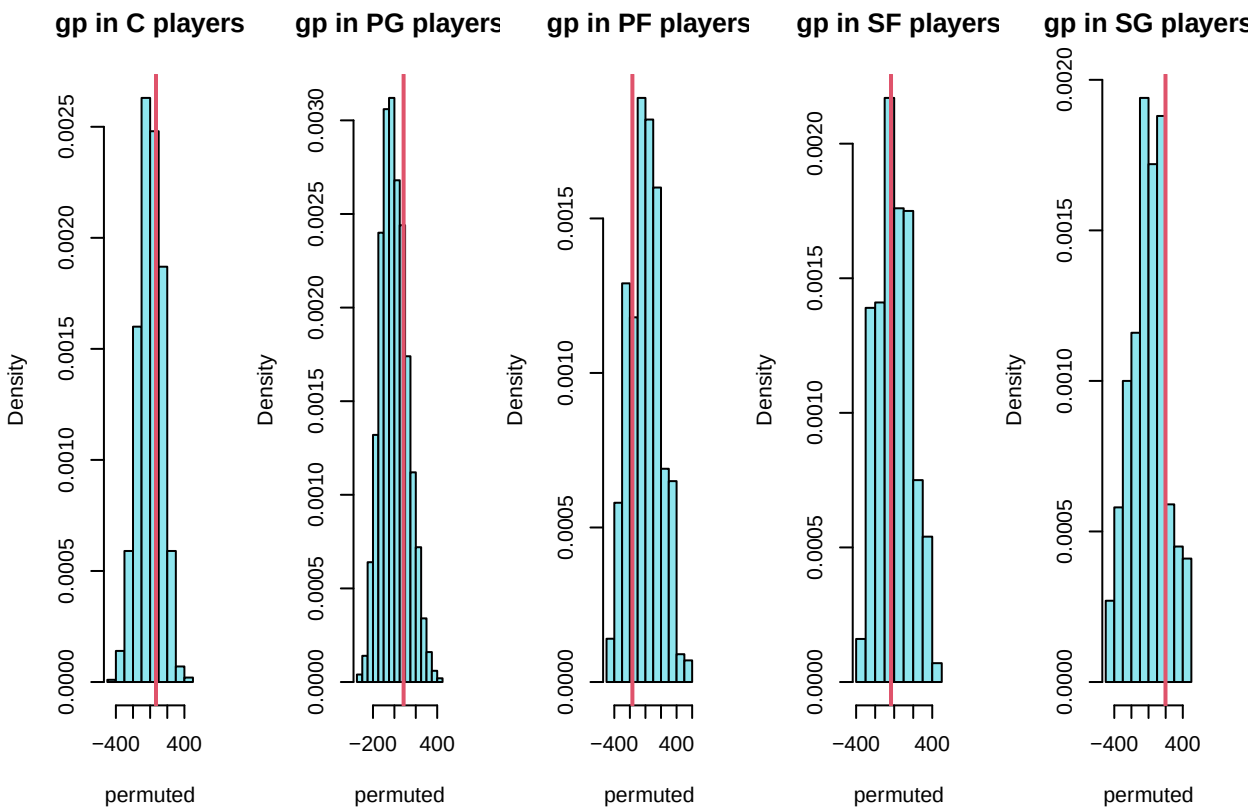
```
set.seed(10)
for(j in c("gp","pctFG","pctFT","minutesTotals","trebTotals","astTotals","pfTotals","ptsTotals")) {
  par(mfrow = c(1,5))
  count <- 1
  for (i in list(C,PG,PF,SF,SG)) {
    permuted <- replicate(1000,permute(i,j))
    observed <- mean_diff(i,j)
    hist(permuted, main = paste(j,"in",c("C","PG","PF","SF","SG")[count],"players",sep=" ") , prob=T, c
    abline(v = observed , col = 2, lwd = 2)
    print(paste0("The observed difference of ",j," is higher than ",mean(permuted<observed)," of the pe
    count <- count + 1
  }
}
```

```
## [1] "The observed difference of gp is higher than 0.676 of the permuted differences for C players"
```

```
## [1] "The observed difference of gp is higher than 0.762 of the permuted differences for PG players"
```

```
## [1] "The observed difference of gp is higher than 0.23 of the permuted differences for PF players"
```

```
## [1] "The observed difference of gp is higher than 0.453 of the permuted differences for SF players"
```



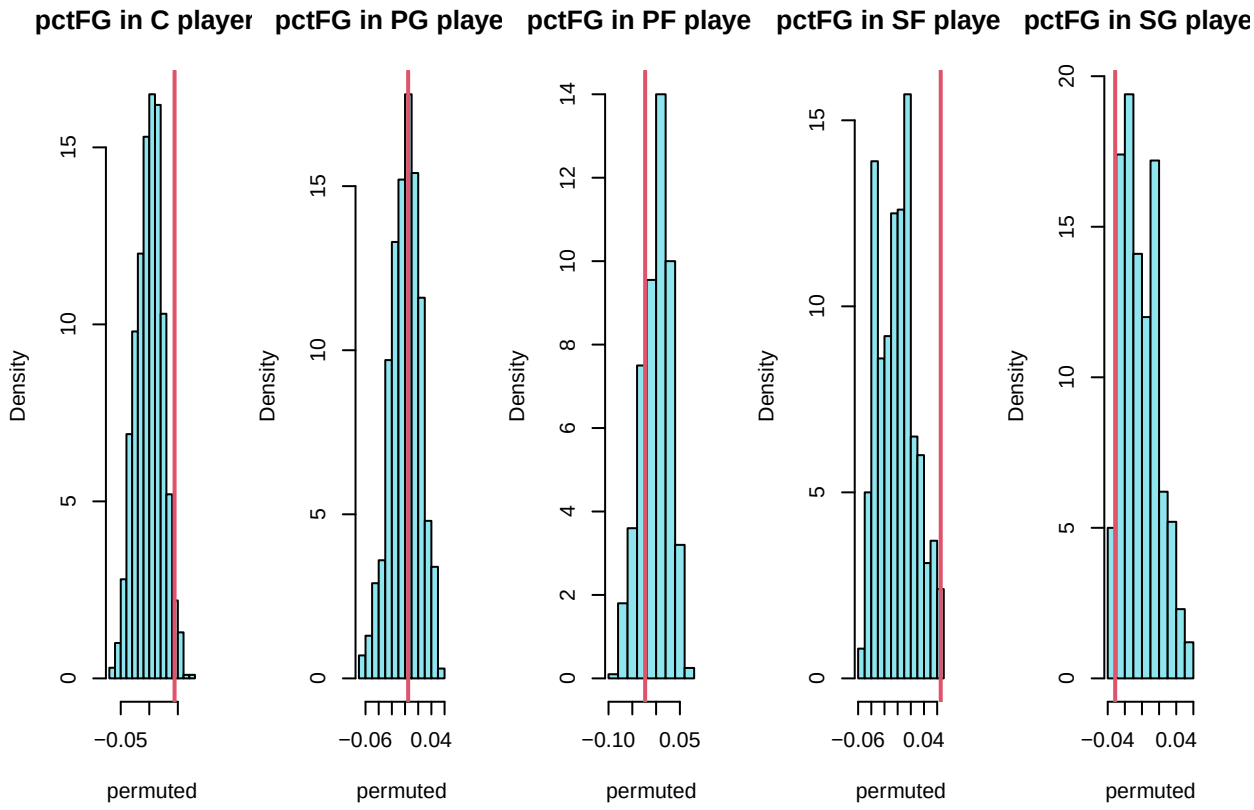

```
## [1] "The observed difference of gp is higher than 0.848 of the permuted differences for SG players"

## [1] "The observed difference of pctFG is higher than 0.973 of the permuted differences for C players"

## [1] "The observed difference of pctFG is higher than 0.555 of the permuted differences for PG players"

## [1] "The observed difference of pctFG is higher than 0.213 of the permuted differences for PF players"

## [1] "The observed difference of pctFG is higher than 0.976 of the permuted differences for SF players"
```



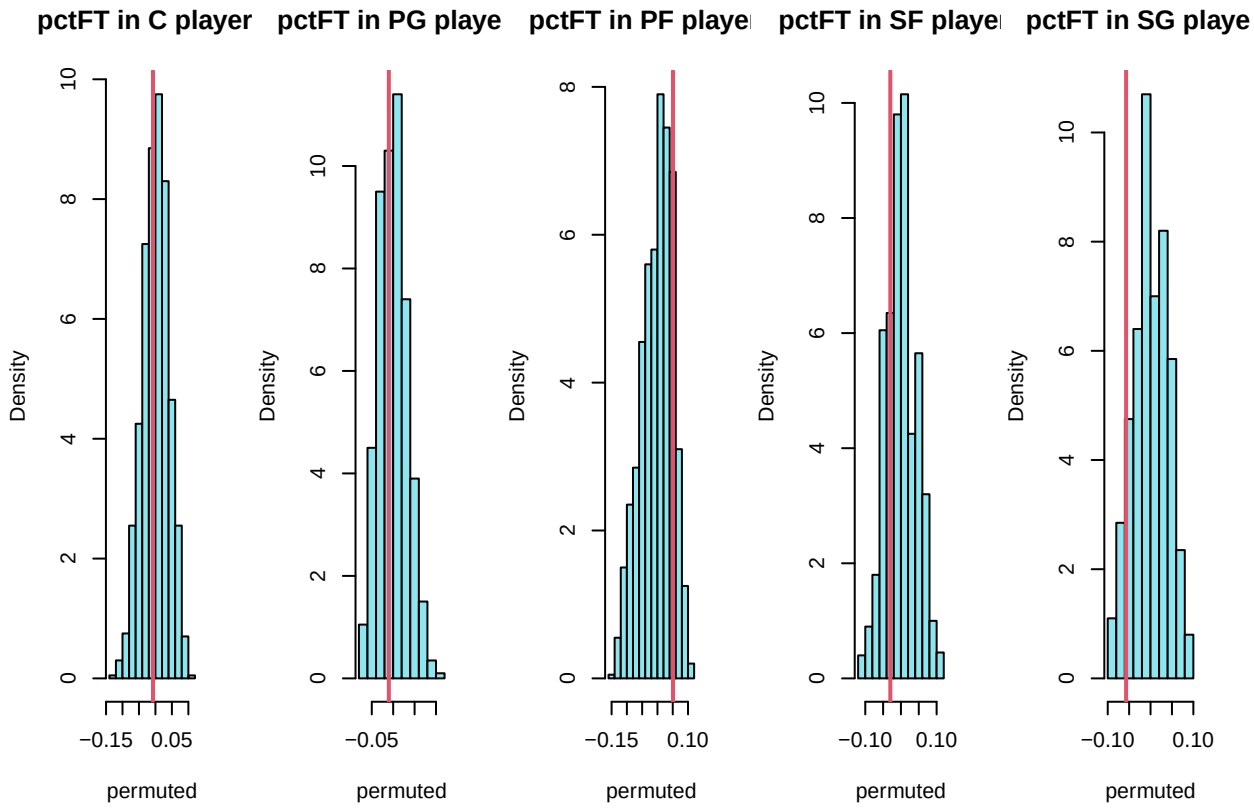

```
## [1] "The observed difference of pctFG is higher than 0.042 of the permuted differences for SG players

## [1] "The observed difference of pctFT is higher than 0.417 of the permuted differences for C players

## [1] "The observed difference of pctFT is higher than 0.392 of the permuted differences for PG players

## [1] "The observed difference of pctFT is higher than 0.849 of the permuted differences for PF players

## [1] "The observed difference of pctFT is higher than 0.272 of the permuted differences for SF players
```



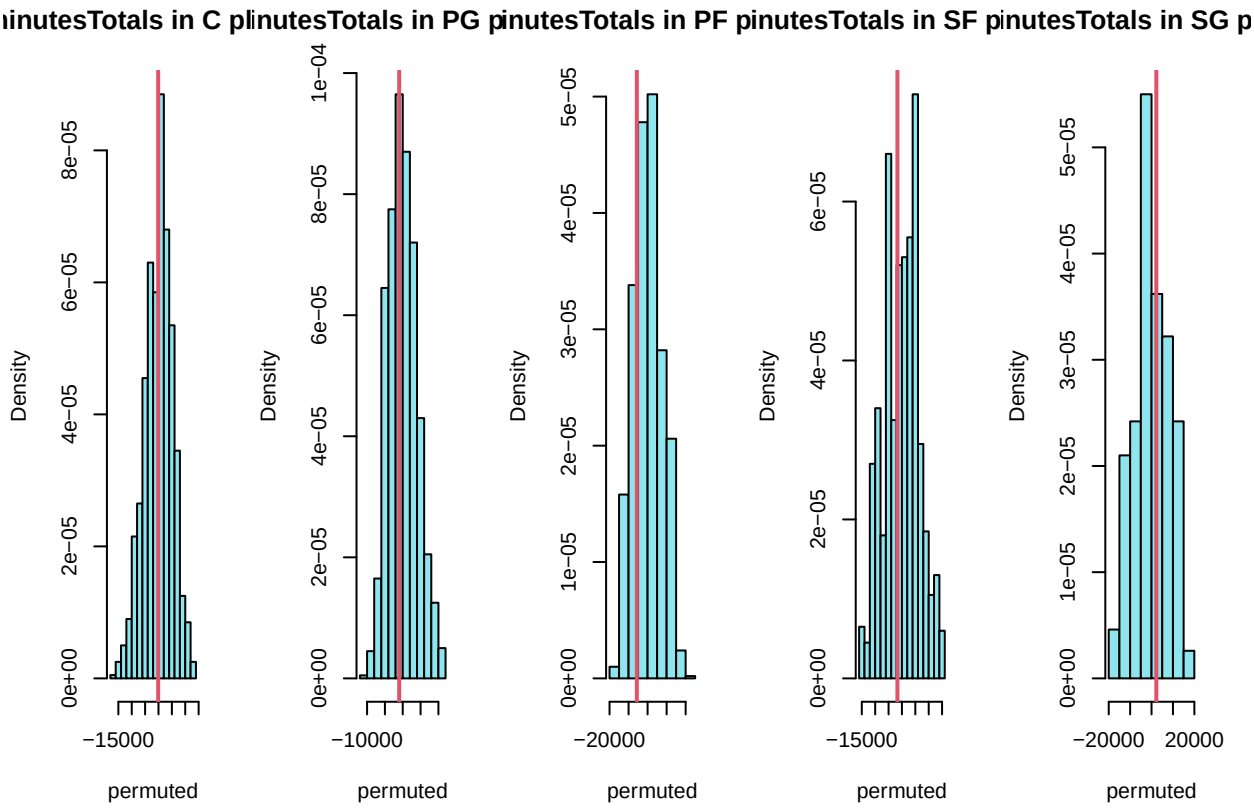

```
## [1] "The observed difference of pctFT is higher than 0.079 of the permuted differences for SG players

## [1] "The observed difference of minutesTotals is higher than 0.458 of the permuted differences for C

## [1] "The observed difference of minutesTotals is higher than 0.421 of the permuted differences for PG

## [1] "The observed difference of minutesTotals is higher than 0.226 of the permuted differences for PF

## [1] "The observed difference of minutesTotals is higher than 0.405 of the permuted differences for SF
```



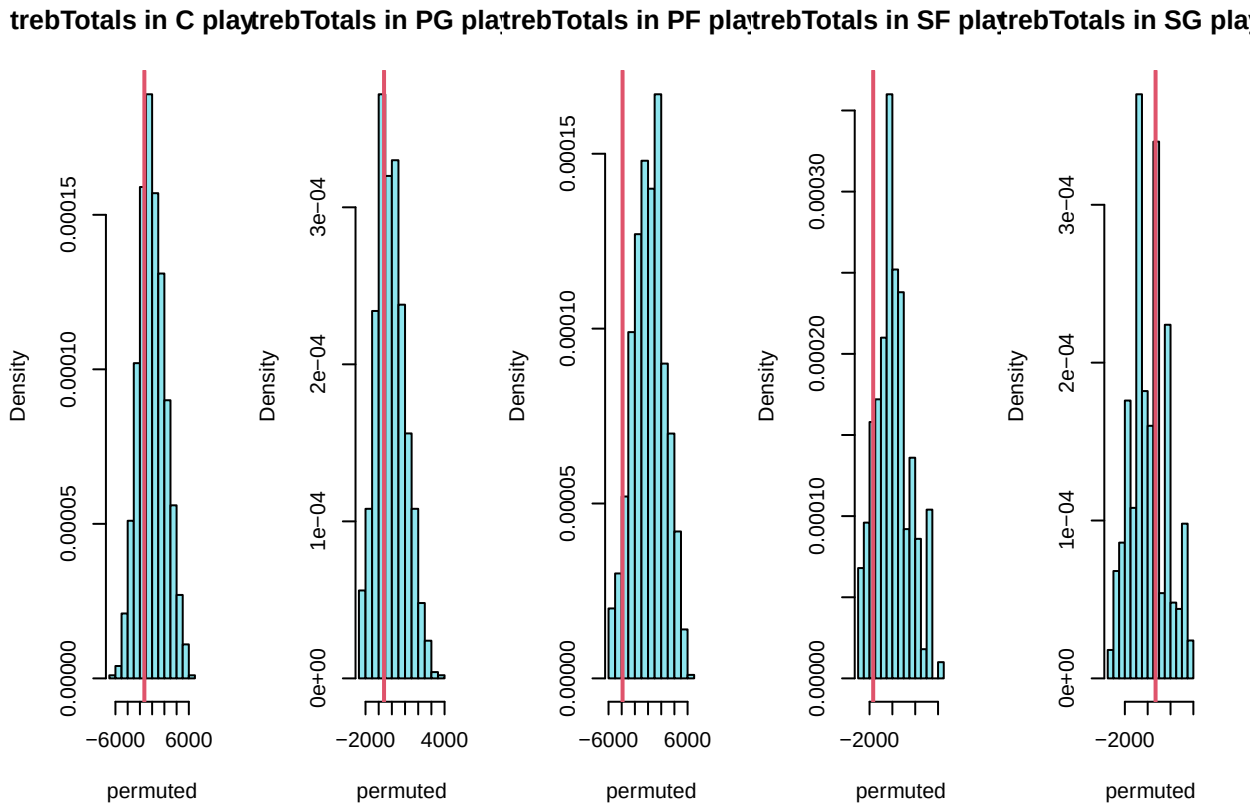

```
## [1] "The observed difference of minutesTotals is higher than 0.592 of the permuted differences for S

## [1] "The observed difference of trebTotals is higher than 0.297 of the permuted differences for C pla

## [1] "The observed difference of trebTotals is higher than 0.351 of the permuted differences for PG pl

## [1] "The observed difference of trebTotals is higher than 0.053 of the permuted differences for PF pl

## [1] "The observed difference of trebTotals is higher than 0.121 of the permuted differences for SF pl
```



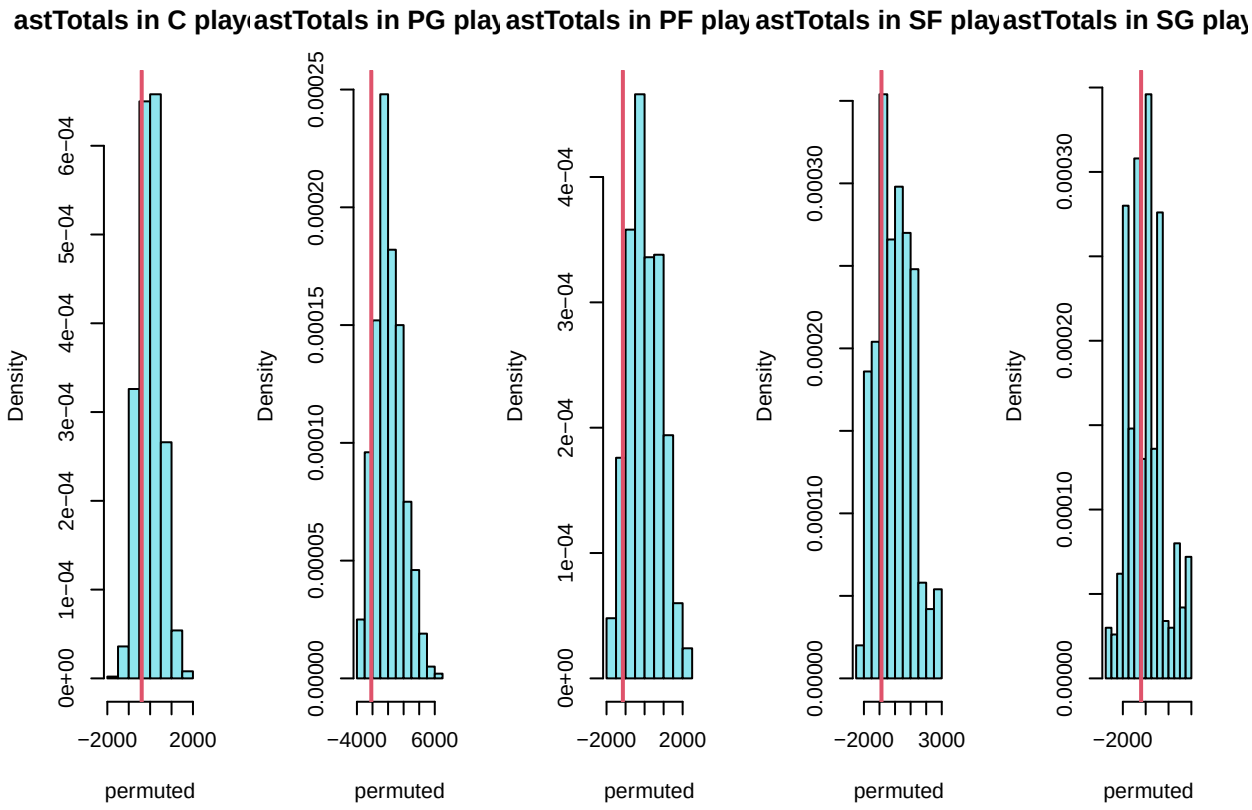

```
## [1] "The observed difference of trebTotals is higher than 0.648 of the permuted differences for SG pl

## [1] "The observed difference of astTotals is higher than 0.248 of the permuted differences for C play

## [1] "The observed difference of astTotals is higher than 0.103 of the permuted differences for PG pla

## [1] "The observed difference of astTotals is higher than 0.073 of the permuted differences for PF pla

## [1] "The observed difference of astTotals is higher than 0.272 of the permuted differences for SF pla
```



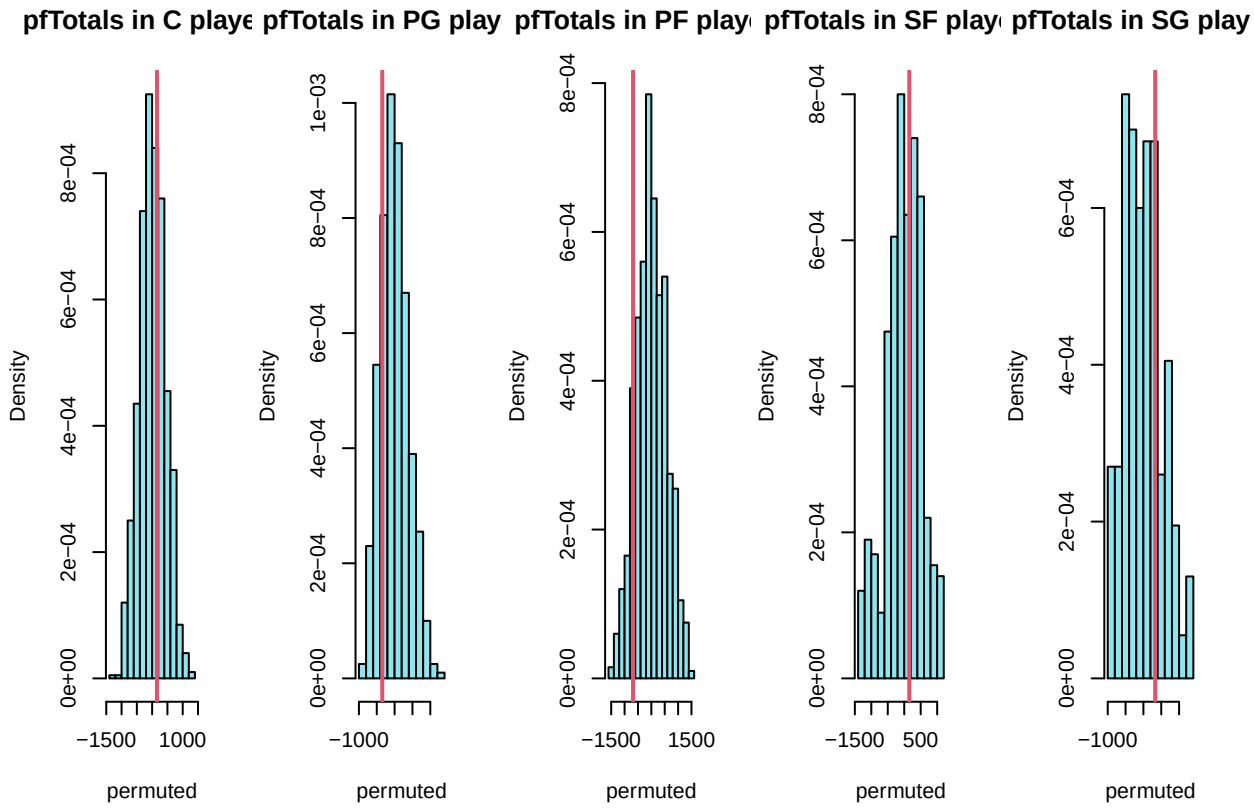

```
## [1] "The observed difference of astTotals is higher than 0.437 of the permuted differences for SG pla

## [1] "The observed difference of pfTotals is higher than 0.624 of the permuted differences for C playe

## [1] "The observed difference of pfTotals is higher than 0.208 of the permuted differences for PG play

## [1] "The observed difference of pfTotals is higher than 0.116 of the permuted differences for PF play

## [1] "The observed difference of pfTotals is higher than 0.598 of the permuted differences for SF play
```



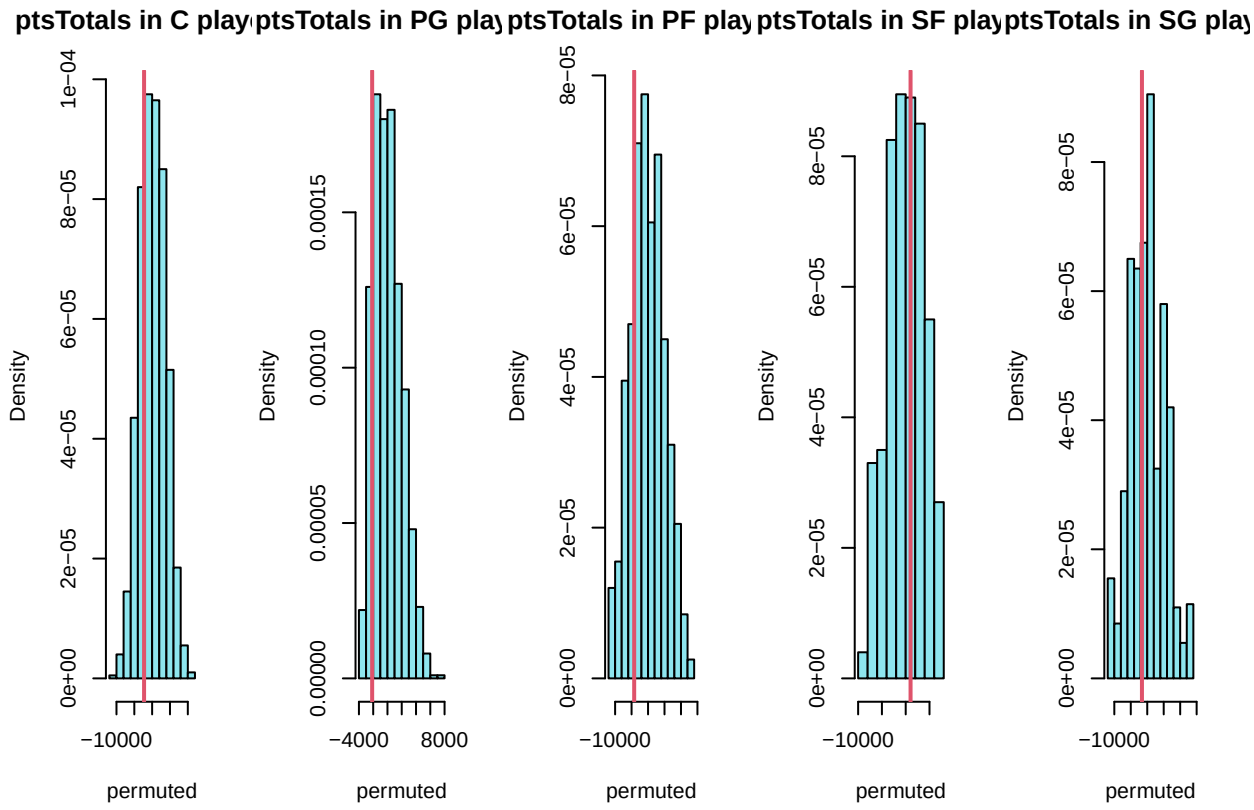

```
## [1] "The observed difference of pfTotals is higher than 0.715 of the permuted differences for SG play

## [1] "The observed difference of ptsTotals is higher than 0.265 of the permuted differences for C play

## [1] "The observed difference of ptsTotals is higher than 0.129 of the permuted differences for PG pla

## [1] "The observed difference of ptsTotals is higher than 0.221 of the permuted differences for PF pla

## [1] "The observed difference of ptsTotals is higher than 0.622 of the permuted differences for SF pla
```




```
## [1] "The observed difference of ptsTotals is higher than 0.381 of the permuted differences for SG pla
```

None of groups has significant difference in gp. Snubbed Center and Short Forward players have a significantly higher pctFG value, snubbed Shoot Guard players have significantly lower value. Snubbed Shoot Guard players have significantly lower value in pctFT. Snubbed Point Forward players have significantly lower value in trebTotals and astTotals. For snubbed players: Center achieve a higher average of field goal rate Shoot Guard have a lower average of field goal rate and free throw rate. Point Forward have a lower average of rebounds and assists.

Method 3: t-test 1. For all players

```
snubbed <- player_data[player_data$Status == 0,]
listed <- player_data[player_data$Status == 1,]
t.test(snubbed$gp,listed$gp,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$gp and listed$gp
## t = 0.59902, df = 28.95, p-value = 0.5538
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -83.64392 152.92725
## sample estimates:
## mean of x mean of y
##  1052.375  1017.733
```

```
t.test(snubbed$pctFG,listed$pctFG,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$pctFG and listed$pctFG
## t = 1.6651, df = 19.365, p-value = 0.112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.005871401  0.051851601
## sample estimates:
## mean of x mean of y
## 0.4988625 0.4758724
```

```
t.test(snubbed$pctFT,listed$pctFT,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$pctFT and listed$pctFT
## t = -1.0576, df = 22.364, p-value = 0.3015
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06998757  0.02268325
## sample estimates:
## mean of x mean of y
## 0.7549964 0.7786485
```

```
t.test(snubbed$minutesTotals,listed$minutesTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$minutesTotals and listed$minutesTotals
## t = -0.6141, df = 41.896, p-value = 0.5425
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4635.604  2472.724
## sample estimates:
## mean of x mean of y
##  34753.00  35834.44
```

```
t.test(snubbed$trebTotals,listed$trebTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$trebTotals and listed$trebTotals
## t = -0.2498, df = 26.706, p-value = 0.8047
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -2439.555  1910.257
## sample estimates:
## mean of x mean of y
##  7935.938  8200.587
```

```
t.test(snubbed$astTotals,listed$astTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$astTotals and listed$astTotals
## t = -2.7653, df = 38.836, p-value = 0.008657
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2475.6089  -383.7861
## sample estimates:
## mean of x mean of y
##  3254.062  4683.760
```

```
t.test(snubbed$pfTotals,listed$pfTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$pfTotals and listed$pfTotals
## t = 0.046341, df = 24.612, p-value = 0.9634
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -395.2910  413.4743
## sample estimates:
## mean of x mean of y
##  2735.625  2726.533
```

```
t.test(snubbed$ptsTotals,listed$ptsTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$ptsTotals and listed$ptsTotals
## t = -1.884, df = 44.772, p-value = 0.06607
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4464.8361   149.3228
## sample estimates:
## mean of x mean of y
##  18701.75  20859.51
```

Seems there's significant difference in assist and total points.

```
t.test(snubbed$astTotals,listed$astTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$astTotals and listed$astTotals
## t = -2.7653, df = 38.836, p-value = 0.004329
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -558.4951
## sample estimates:
## mean of x mean of y
##  3254.062  4683.760
```

```
t.test(snubbed$ptsTotals,listed$ptsTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  snubbed$ptsTotals and listed$ptsTotals
## t = -1.884, df = 44.772, p-value = 0.03303
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -234.1004
## sample estimates:
## mean of x mean of y
##  18701.75  20859.51
```

The listed players have significantly higher numbers of assists and points than snubbed players.

What if for different positions?

```
C_0 <- C[C$Status==0,]
C_1 <- C[C$Status==1,]
t.test(C_0$gp,C_1$gp,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$gp and C_1$gp
## t = 0.71327, df = 20.424, p-value = 0.4837
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -131.2411  267.9078
## sample estimates:
## mean of x mean of y
##  1089.333  1021.000
```

```
t.test(C_0$pctFG,C_1$pctFG,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$pctFG and C_1$pctFG
```

```
## t = 2.2008, df = 9.7798, p-value = 0.05296
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0006867615  0.0891855393
## sample estimates:
## mean of x mean of y
## 0.5402193 0.4959699
```

```
t.test(C_0$pctFT,C_1$pctFT,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$pctFT and C_1$pctFT
## t = -0.19681, df = 9.6542, p-value = 0.8481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09064201  0.07599467
## sample estimates:
## mean of x mean of y
## 0.6858203 0.6931440
```

```
t.test(C_0$minutesTotals,C_1$minutesTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$minutesTotals and C_1$minutesTotals
## t = -0.035728, df = 21.783, p-value = 0.9718
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7312.766  7065.210
## sample estimates:
## mean of x mean of y
##  36221.17  36344.94
```

```
t.test(C_0$trebTotals,C_1$trebTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$trebTotals and C_1$trebTotals
## t = -0.84365, df = 19.978, p-value = 0.4089
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4413.073  1871.517
## sample estimates:
## mean of x mean of y
##  11857.67  13128.44
```

```r
t.test(C_0$astTotals,C_1$astTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$astTotals and C_1$astTotals
## t = -0.79781, df = 10.085, p-value = 0.4434
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1493.5415   705.3193
## sample estimates:
## mean of x mean of y
##  2357.333  2751.444
```

```r
t.test(C_0$pfTotals,C_1$pfTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$pfTotals and C_1$pfTotals
## t = 0.53898, df = 19.887, p-value = 0.5959
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -453.0736  768.6291
## sample estimates:
## mean of x mean of y
##  3237.000  3079.222
```

```r
t.test(C_0$ptsTotals,C_1$ptsTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$ptsTotals and C_1$ptsTotals
## t = -0.88934, df = 20.588, p-value = 0.3841
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7548.040  3029.929
## sample estimates:
## mean of x mean of y
##  17935.50  20194.56
```

```r
t.test(C_0$pctFG,C_1$pctFG,"greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  C_0$pctFG and C_1$pctFG
## t = 2.2008, df = 9.7798, p-value = 0.02648
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  0.007724536         Inf
## sample estimates:
## mean of x mean of y
## 0.5402193 0.4959699
```

Snubbed Center players seem to have greater pctFG, at 5% level.

```
PF_0 <- PF[PF$Status==0,]
PF_1 <- PF[PF$Status==1,]
t.test(PF_0$gp,PF_1$gp,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$gp and PF_1$gp
## t = -1.2558, df = 7.0697, p-value = 0.2491
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -482.996  147.496
## sample estimates:
## mean of x mean of y
##  890.3333 1058.0833
```

```
t.test(PF_0$pctFG,PF_1$pctFG,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$pctFG and PF_1$pctFG
## t = -0.97762, df = 4.5284, p-value = 0.3776
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08548891  0.03945258
## sample estimates:
## mean of x mean of y
## 0.4738070 0.4968252
```

```
t.test(PF_0$pctFT,PF_1$pctFT,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$pctFT and PF_1$pctFT
## t = 1.7777, df = 11.663, p-value = 0.1015
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01165051  0.11315001
## sample estimates:
## mean of x mean of y
## 0.8196300 0.7688803
```

```
t.test(PF_0$minutesTotals,PF_1$minutesTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$minutesTotals and PF_1$minutesTotals
## t = -1.176, df = 7.3117, p-value = 0.2765
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16914.515   5613.515
## sample estimates:
## mean of x mean of y
##  30687.67  36338.17
```

```
t.test(PF_0$trebTotals,PF_1$trebTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$trebTotals and PF_1$trebTotals
## t = -2.4679, df = 5.7266, p-value = 0.0505
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7754.07014    11.90347
## sample estimates:
## mean of x mean of y
##  6958.333 10829.417
```

```
t.test(PF_0$astTotals,PF_1$astTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$astTotals and PF_1$astTotals
## t = -2.4771, df = 12.105, p-value = 0.02896
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2144.5865  -138.4135
## sample estimates:
## mean of x mean of y
##    2143.0    3284.5
```

```
t.test(PF_0$pfTotals,PF_1$pfTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$pfTotals and PF_1$pfTotals
## t = -2.1205, df = 10.86, p-value = 0.05783
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -1393.03601    27.03601
## sample estimates:
## mean of x mean of y
##  2217.333  2900.333
```

```
t.test(PF_0$ptsTotals,PF_1$ptsTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$ptsTotals and PF_1$ptsTotals
## t = -1.4596, df = 11.305, p-value = 0.1716
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10534.740   2116.907
## sample estimates:
## mean of x mean of y
##  17420.67  21629.58
```

```
t.test(PF_0$trebTotals,PF_1$trebTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$trebTotals and PF_1$trebTotals
## t = -2.4679, df = 5.7266, p-value = 0.02525
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -796.8266
## sample estimates:
## mean of x mean of y
##  6958.333 10829.417
```

```
t.test(PF_0$astTotals,PF_1$astTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  PF_0$astTotals and PF_1$astTotals
## t = -2.4771, df = 12.105, p-value = 0.01448
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -320.7709
## sample estimates:
## mean of x mean of y
##    2143.0    3284.5
```

```
t.test(PF_0$pfTotals,PF_1$pfTotals,"less")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  PF_0$pfTotals and PF_1$pfTotals
## t = -2.1205, df = 10.86, p-value = 0.02892
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -103.8764
## sample estimates:
## mean of x mean of y
##  2217.333  2900.333
```

Snubbed Power Forward players seem to have lower assist, pf and rebounds, on 5% level.

```
PG_0 <- PG[PG$Status==0,]
PG_1 <- PG[PG$Status==1,]
t.test(PG_0$gp,PG_1$gp,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$gp and PG_1$gp
## t = 0.99079, df = 3.8549, p-value = 0.3798
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -158.5513  330.4846
## sample estimates:
## mean of x mean of y
##  1107.667  1021.700
```

```
t.test(PG_0$pctFG,PG_1$pctFG,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$pctFG and PG_1$pctFG
## t = 0.18583, df = 2.4543, p-value = 0.8668
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08746978  0.09692868
## sample estimates:
## mean of x mean of y
## 0.4632710 0.4585416
```

```
t.test(PG_0$pctFT,PG_1$pctFT,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$pctFT and PG_1$pctFT
## t = -0.21975, df = 2.2449, p-value = 0.8445
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1895990  0.1692873
```

```
## sample estimates:
## mean of x mean of y
## 0.8048003 0.8149562
```

```
t.test(PG_0$minutesTotals,PG_1$minutesTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$minutesTotals and PG_1$minutesTotals
## t = -0.45437, df = 7.1271, p-value = 0.6631
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6126.312  4145.412
## sample estimates:
## mean of x mean of y
##  35354.00  36344.45
```

```
t.test(PG_0$trebTotals,PG_1$trebTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$trebTotals and PG_1$trebTotals
## t = -0.55374, df = 2.658, p-value = 0.6229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4270.486  3082.420
## sample estimates:
## mean of x mean of y
##  4185.667  4779.700
```

```
t.test(PG_0$astTotals,PG_1$astTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$astTotals and PG_1$astTotals
## t = -2.0417, df = 5.6392, p-value = 0.09026
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4778.4717   468.3383
## sample estimates:
## mean of x mean of y
##  5641.333  7796.400
```

```
t.test(PG_0$pfTotals,PG_1$pfTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  PG_0$pfTotals and PG_1$pfTotals
## t = -1.9589, df = 15.121, p-value = 0.06884
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -722.70991   30.24324
## sample estimates:
## mean of x mean of y
##  2201.667  2547.900
```

```
t.test(PG_0$ptsTotals,PG_1$ptsTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$ptsTotals and PG_1$ptsTotals
## t = -1.6784, df = 3.9526, p-value = 0.1694
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5709.233  1419.867
## sample estimates:
## mean of x mean of y
##  17470.67  19615.35
```

```
t.test(PG_0$astTotals,PG_1$astTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$astTotals and PG_1$astTotals
## t = -2.0417, df = 5.6392, p-value = 0.04513
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -80.2172
## sample estimates:
## mean of x mean of y
##  5641.333  7796.400
```

```
t.test(PG_0$pfTotals,PG_1$pfTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  PG_0$pfTotals and PG_1$pfTotals
## t = -1.9589, df = 15.121, p-value = 0.03442
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -36.54019
## sample estimates:
## mean of x mean of y
##  2201.667  2547.900
```

Snubbed Point Guard players seem to have lower assists and pf, at 5% level.

```r
SF_0 <- SF[SF$Status==0,]
SF_1 <- SF[SF$Status==1,]
t.test(SF_0$gp,SF_1$gp,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$gp and SF_1$gp
## t = -0.41956, df = 8.8419, p-value = 0.6848
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -213.6321  146.9398
## sample estimates:
## mean of x mean of y
##  914.5000  947.8462
```

```r
t.test(SF_0$pctFG,SF_1$pctFG,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$pctFG and SF_1$pctFG
## t = 4.6695, df = 2.5446, p-value = 0.02647
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01595925 0.11486367
## sample estimates:
## mean of x mean of y
## 0.5292380 0.4638265
```

```r
t.test(SF_0$pctFT,SF_1$pctFT,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$pctFT and SF_1$pctFT
## t = -0.63338, df = 1.2691, p-value = 0.6213
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3969709  0.3374589
## sample estimates:
## mean of x mean of y
## 0.7738575 0.8036135
```

```r
t.test(SF_0$minutesTotals,SF_1$minutesTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$minutesTotals and SF_1$minutesTotals
## t = -0.48907, df = 4.0527, p-value = 0.6501
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11191.690   7824.767
## sample estimates:
## mean of x mean of y
##  31784.00  33467.46
```

```
t.test(SF_0$trebTotals,SF_1$trebTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$trebTotals and SF_1$trebTotals
## t = -2.8487, df = 12.775, p-value = 0.01389
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2954.9148  -403.4698
## sample estimates:
## mean of x mean of y
##  5257.500  6936.692
```

```
t.test(SF_0$astTotals,SF_1$astTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$astTotals and SF_1$astTotals
## t = -2.0412, df = 12.036, p-value = 0.06379
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1803.68379    58.52994
## sample estimates:
## mean of x mean of y
##  2846.500  3719.077
```

```
t.test(SF_0$pfTotals,SF_1$pfTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$pfTotals and SF_1$pfTotals
## t = 0.58434, df = 5.1527, p-value = 0.5836
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -516.3048  823.6125
## sample estimates:
## mean of x mean of y
##  2717.500  2563.846
```

```
t.test(SF_0$ptsTotals,SF_1$ptsTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$ptsTotals and SF_1$ptsTotals
## t = 0.44227, df = 2.8515, p-value = 0.6897
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6559.887  8605.887
## sample estimates:
## mean of x mean of y
##     21416     20393
```

```
t.test(SF_0$pctFG,SF_1$pctFG,"greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$pctFG and SF_1$pctFG
## t = 4.6695, df = 2.5446, p-value = 0.01323
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.02989485        Inf
## sample estimates:
## mean of x mean of y
## 0.5292380 0.4638265
```

```
t.test(SF_0$trebTotals,SF_1$trebTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$trebTotals and SF_1$trebTotals
## t = -2.8487, df = 12.775, p-value = 0.006946
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -633.8994
## sample estimates:
## mean of x mean of y
##  5257.500  6936.692
```

```
t.test(SF_0$astTotals,SF_1$astTotals,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  SF_0$astTotals and SF_1$astTotals
## t = -2.0412, df = 12.036, p-value = 0.03189
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -110.8615
## sample estimates:
## mean of x mean of y
##  2846.500  3719.077
```

Snubbed Short Forward players seem to have higher pctFG (5%) and lower assist (5%). And lower rebounds at 1% level.

```
SG_0 <- SG[SG$Status==0,]
SG_1 <- SG[SG$Status==1,]
t.test(SG_0$gp,SG_1$gp,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$gp and SG_1$gp
## t = 0.63733, df = 1.1251, p-value = 0.6289
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2848.307  3244.140
## sample estimates:
## mean of x mean of y
##  1239.500  1041.583
```

```
t.test(SG_0$pctFG,SG_1$pctFG,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$pctFG and SG_1$pctFG
## t = -4.0315, df = 11.004, p-value = 0.001975
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04841929 -0.01422221
## sample estimates:
## mean of x mean of y
## 0.4353870 0.4667077
```

```
t.test(SG_0$pctFT,SG_1$pctFT,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$pctFT and SG_1$pctFT
## t = -1.9347, df = 1.71, p-value = 0.2138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.20716584  0.09294901
## sample estimates:
## mean of x mean of y
## 0.7720070 0.8291154
```

```
t.test(SG_0$minutesTotals,SG_1$minutesTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  SG_0$minutesTotals and SG_1$minutesTotals
## t = 0.26514, df = 1.3218, p-value = 0.8263
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -59257.37  63727.04
## sample estimates:
## mean of x mean of y
##  38514.00  36279.17
```

```
t.test(SG_0$trebTotals,SG_1$trebTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$trebTotals and SG_1$trebTotals
## t = 0.7784, df = 2.9981, p-value = 0.4931
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2133.081  3513.748
## sample estimates:
## mean of x mean of y
##  5941.000  5250.667
```

```
t.test(SG_0$astTotals,SG_1$astTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$astTotals and SG_1$astTotals
## t = -0.57411, df = 11.215, p-value = 0.5772
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1936.334  1133.668
## sample estimates:
## mean of x mean of y
##  4437.500  4838.833
```

```
t.test(SG_0$pfTotals,SG_1$pfTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$pfTotals and SG_1$pfTotals
## t = 0.28072, df = 1.0337, p-value = 0.8246
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13516.48  14177.15
## sample estimates:
## mean of x mean of y
##  2828.000  2497.667
```

```
t.test(SG_0$ptsTotals,SG_1$ptsTotals,"two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$ptsTotals and SG_1$ptsTotals
## t = -0.38076, df = 1.7443, p-value = 0.7447
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22633.07  19410.40
## sample estimates:
## mean of x mean of y
##  22054.50  23665.83
```

```
t.test(SG_0$pctFG,SG_1$pctFG,"less")
```

```
##
##  Welch Two Sample t-test
##
## data:  SG_0$pctFG and SG_1$pctFG
## t = -4.0315, df = 11.004, p-value = 0.0009877
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01736911
## sample estimates:
## mean of x mean of y
## 0.4353870 0.4667077
```

Snubbed Score Guard players seem to have lower pctFG, at 1% level.