How to Win Big Brother: A Statistical Analysis

Alissa Trujillo

2022-05-17

Introduction

Big Brother is a television game show that involves a number of strangers being cut off from the outside world and living in a house together. They have no contact with their families, access to the news, or most of the luxuries of everyday life. They compete in competitions every week to determine who is the Head of Household, who must nominate two houseguests for eviction, and who holds the Power of Veto, which can save one of the nominees on the block. For the first phase of the game, the players who are evicted are sent home. In the second phase of the game, the players who are evicted become members of the jury. They remain sequestered from the outside world and cast votes at the end of the season for the winner.

What Makes a Good Player?

I have collected data on each of the 370 houseguests who have entered the Big Brother house over the 23 seasons of the show's run. This data includes competition wins, demographic information, and how well the contestants fared in the game. The purpose of this paper will be to analyze which factors contribute to a houseguest being a good Big Brother player, making it far in the game, and ultimately becoming a winner. I will also be taking a look at some commonly-held beliefs by fans, such as the "First to Enter Curse" and that being the first HOH puts a player on the path to success.

The Dataset

The dataset I have compiled takes information from three separate data sets I collected. The first is contestant_data, which was compiled by github user vdixon3 ("Big Brother's Diversity Problem" (2019)). This data set boasts 49 variables containing information about each of the 370 houseguests to ever set foot the Big Brother house. The second data set, additional_data is information I compiled myself using online sources regarding competition wins and miscellaneous information I needed to test my hypotheses. This data set includes 7 variables, 3 of which are identical to contestant_data, so they are easily able to combine together. The third data set, diversity_data includes season-specific information about diversity and demographics. This data had to be transformed from season data to contestant data. For integrity, the data from season 1 had to be removed due to the format of the show being completely different and most of the values for important variables being N/A. The final data set contains 22 variables regarding the 360 houseguests from season 2 and beyond.

```
##    first       last season_code age gender ethnicity   poc season_diversity
## 1   Will      Kirby       bbus2  28   male     white FALSE             0.17
## 2 Nicole Schaffrich       bbus2  31 female     white FALSE             0.17
## 3 Monica     Bailey       bbus2  40 female     black  TRUE             0.17
## 4  Hardy  Ames-Hill       bbus2  31   male     white FALSE             0.17
## 5  Bunky     Miller       bbus2  36   male     white FALSE             0.17
## 6 Krista    Stegall       bbus2  28 female     white FALSE             0.17
```

```
##       lgbt appearance first_in_house comps hoh veto first_hoh nom afp
## 1 non-lgbt          1          FALSE     0   0   na     FALSE   4  na
## 2 non-lgbt          1           TRUE     1   1   na     FALSE   2  na
## 3 non-lgbt          1          FALSE     1   1   na     FALSE   2  na
## 4 non-lgbt          1          FALSE     3   3   na     FALSE   1  na
## 5      gay          1          FALSE     0   0   na     FALSE   1  na
## 6 non-lgbt          1          FALSE     1   1   na     FALSE   2  na
##   jurymember placement final2 runnerup winner
## 1       TRUE         1   TRUE    FALSE   TRUE
## 2       TRUE         2   TRUE     TRUE  FALSE
## 3       TRUE         3  FALSE    FALSE  FALSE
## 4       TRUE         4  FALSE    FALSE  FALSE
## 5       TRUE         5  FALSE    FALSE  FALSE
## 6       TRUE         6  FALSE    FALSE  FALSE
```

Are The Rumors True?

I am going to begin my analysis by taking a look at a couple of theories that are floated around by fans of the show.

First To Enter Curse

It is widely believed that the first houseguest to enter the house each season is doomed to do poorly in the game. It is true that no houseguest has ever won the game after stepping through the doors first, but is it statistically significant?

```
cor.test(as.integer(houseguests_df$first_in_house), as.integer(houseguests_df$winner))
```

```
##
##  Pearson's product-moment correlation
##
## data:  as.integer(houseguests_df$first_in_house) and as.integer(houseguests_df$winner)
## t = -0.59899, df = 358, p-value = 0.5496
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13456352  0.07195536
## sample estimates:
##         cor
## -0.03164179
```

The correlation between being the first houseguest to enter the house and being the winner of the game is -0.03, very close to 0 with quite a large confidence interval. The p-value is 0.55, meaning that the correlation is not significant. Only 1 in 16 houseguests can be the first to walk through the door, and only 1 in 16 houseguests will go on to win the game. It is more likely that the two events have never co-occured purely by chance rather than something nefarious occuring.

```
##
##  Pearson's product-moment correlation
##
## data:  as.integer(houseguests_df$first_in_house) and houseguests_df$placement
## t = -0.86851, df = 358, p-value = 0.3857
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.14851208  0.05778154
## sample estimates:
##         cor
## -0.04585414
```

The correlation between a houseguest's placement in the game and whether or not they entered the house first is -0.05, which is once again very close to 0 with a wide confidence interval. Since placement decreases as a player gets farther in the game, this negative correlation would denote a very small correlation between entering the house first and lasting longer in the game. This is once again, not a significant statistic, and is more than likely just due to happenstance.

First HOH

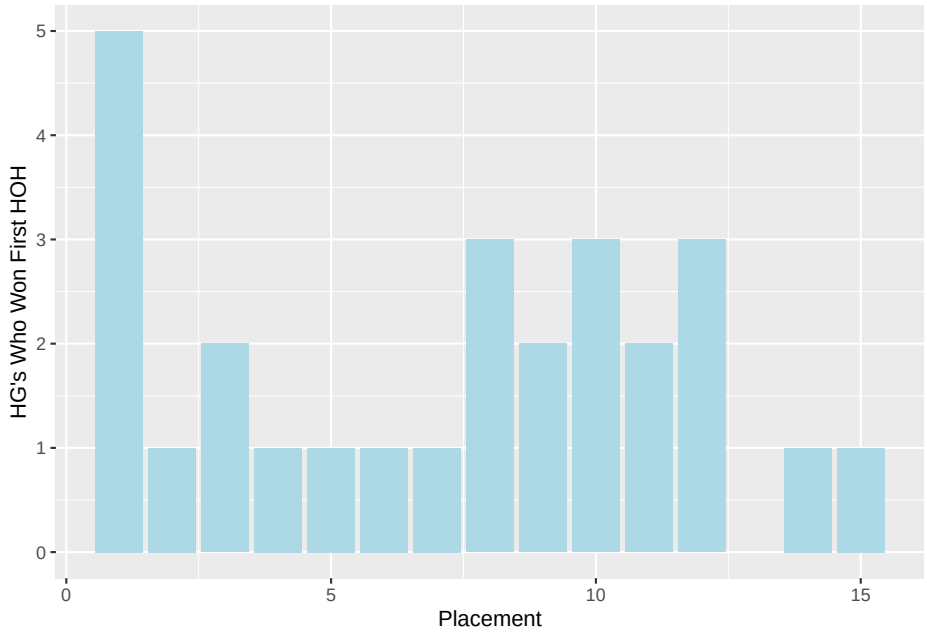
The first HOH is the first semblance of power that is earned throughout the show. This allows the winner to make important relationships and form alliances early on. This is widely believed to be a useful tool for houseguests to establish themselves in the game.

```
##
##  Pearson's product-moment correlation
##
## data:  as.integer(houseguests_df$first_hoh) and as.integer(houseguests_df$winner)
## t = 2.4739, df = 358, p-value = 0.01383
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02664027 0.22992584
## sample estimates:
##       cor
## 0.1296451
```

The correlation between winning the first Head of Household and winning the show is 0.13. This statistic is significant at a p=0.05 level, with a p-value of p=0.01. This bolsters the idea that, with everything else held constant, winning the first competition is correlated positively with winning the game. This supports the hypothesis that forging those early relationships is a really big asset to have as a player.

```
##
##  Pearson's product-moment correlation
##
## data:  as.integer(houseguests_df$first_hoh) and houseguests_df$placement
## t = -0.77388, df = 358, p-value = 0.4395
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.14362199  0.06276008
## sample estimates:
##         cor
## -0.04086684
```

There is a small negative correlation, -0.04, between winning the first HOH competition and a houseguest's final placement in the game. This statistic is not significant, with a p-value of 0.44. This means that winning the first competition is correlated with winning the game, but not with a player's overall final placement. This may be due to the fact that winning so early can make a player look more threatening, so players may be likely to take shots at getting them out early, evening out their final placement.

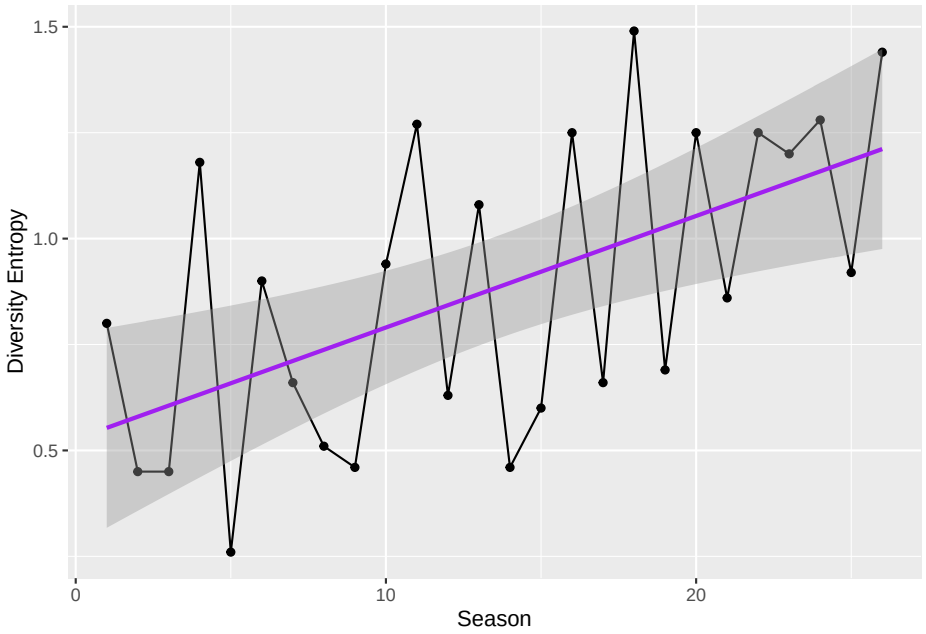


Looking at this graph, we can see that players who win the first HOH tend to go on to win the game, or are voted out right around the beginning of jury (roughly after 5-6 houseguests have been evicted in a 16 HG season, centering around the 10th place mark on this graph). This displays that houseguests who win that very first competition typically fall into two different categories: those that use it to form strong relationships early and set themselves up for success, or those who play too hard too early and get themselves voted out early.

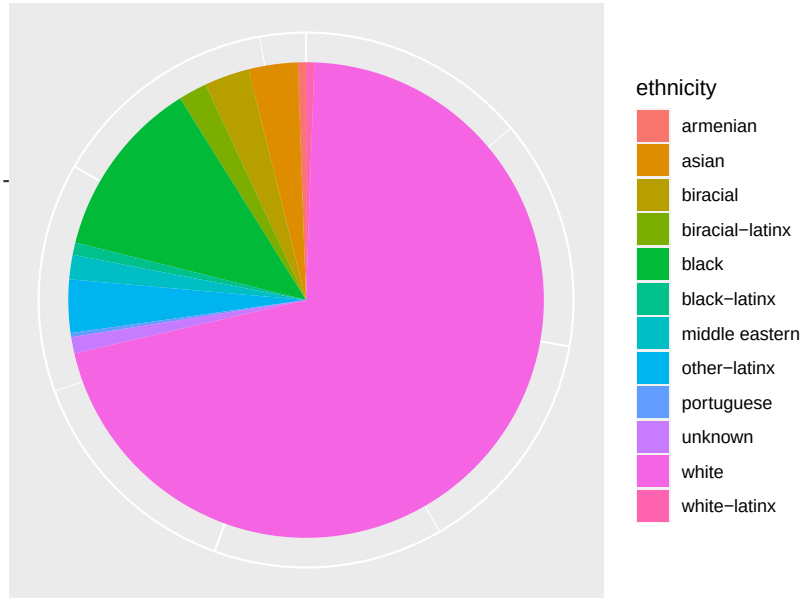
Looking At Diversity

As the seasons have gone on, there has been an increasing amount of diversity in the cast. The houseguests that have been recruited have been more well-rounded and robust characters versus the archetypes of early reality television. While diversity varies season to season, the general trend is that it has increased throughout the show's run.

```
## `geom_smooth()` using formula 'y ~ x'
```



Diversity entropy is defined as how minorities are represented versus their actual makeup of American society. A value below 1 indicates that white players are over represented, and a value above 1 means that as a whole, people of color are represented more prominently than their presence in American communities ("Big Brother's Diversity Problem" (2019)).



Of the 360 houseguests that have competed on the show, 255 have identified as Caucasian. Only 29% of houseguests to ever enter the big brother house consider themselves to be something other than white. While we are taking steps in the right direction, and have just finished one of our most diverse seasons over (resulting in a final 7 comprised entirely of people of color, a black winner, and a black runner up), there is still so much we can do as a society to embrace fully representative television.

Predicting Game Performance

So what characteristics are indicative of a player performing well in the game? We will first analyze the relationship with final game standings, demographic information, and game statistics. Since we have multiple variables we would like to measure against placement, we will be creating a multiple regression model.

The variables that I am predicting to be key indicators of game success are a contestant's competition wins, age, gender, and finally, the interaction between ethnicity and the season's diversity. I am predicting that people of color perform better on seasons where diversity is high. Season diversity is meaningless on its own in this model, as every person in the same season will have the same diversity quotient.

```
##
## Call:
## lm(formula = placement ~ gender + age + poc + poc * season_diversity +
##     comps, data = houseguests_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6342 -2.3147  0.1119  2.2901  7.7388
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               11.31250    0.81556  13.871  < 2e-16 ***
## gendermale                 1.03843    0.35776   2.903  0.00393 **
## age                       -0.09141    0.02337  -3.911  0.00011 ***
## pocTRUE                    2.77765    1.06710   2.603  0.00963 **
## season_diversity           4.95703    1.75475   2.825  0.00500 **
## comps                     -1.43566    0.09167 -15.661  < 2e-16 ***
## pocTRUE:season_diversity -9.05862    3.15315  -2.873  0.00431 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 353 degrees of freedom
## Multiple R-squared:  0.4262, Adjusted R-squared:  0.4164
## F-statistic: 43.69 on 6 and 353 DF,  p-value: < 2.2e-16
```

It is important to remember that when looking at this data, placement goes down as performance goes up (first place is better than second, etc.). So, negative coefficients denote better performance in the game.

The first variable to look at is gender. There is a positive correlation between being male and placement in the game, meaning that women tend to make it farther in the game (significant at p=0.004), with all other variables held constant. This seems surprising as a huge chunk of the winners are men, however it makes sense when you consider that many of the players who are evicted early for playing too hard are also men.

Age has a negative correlation with game performance, meaning that as age increases, placement goes down. This means lasting longer in the game is positively correlated with age (significant at p=0.0001). This is also surprising as younger contestants typically run the game, but upon reflection it makes sense that older houseguests are less likely to be voted out immediately. There are also very few "older" houseguests, most are around 20-30, so the few in their 40's+ are probably outliers and do not weigh heavily on this calculation.

Winning competitions has a high negative correlation with placement in the game. This indicates that those who perform better in competitions last longer in the game (significant at p=0.000). This makes sense, as winning competitions not only boosts your resume in the game, but it allows you to hold power to save yourself or your alliance members each week. This works two-fold, as it allows you to strengthen your relationships with others and prove your worth, but also gives a player concrete powers they can use to

actually protect themselves. I was curious whether this effect would be overshadowed by the fact that this creates a target for the person who is consistently winning, but that seems to not be true.

Identifying as a person of color is correlated positively with placement, meaning that they do not tend to last as long as those who identify as white (significant at p=0.01). This is not surprising, as the early seasons did not provide much diversity and relied heavily on type-casting that created barriers for their success.

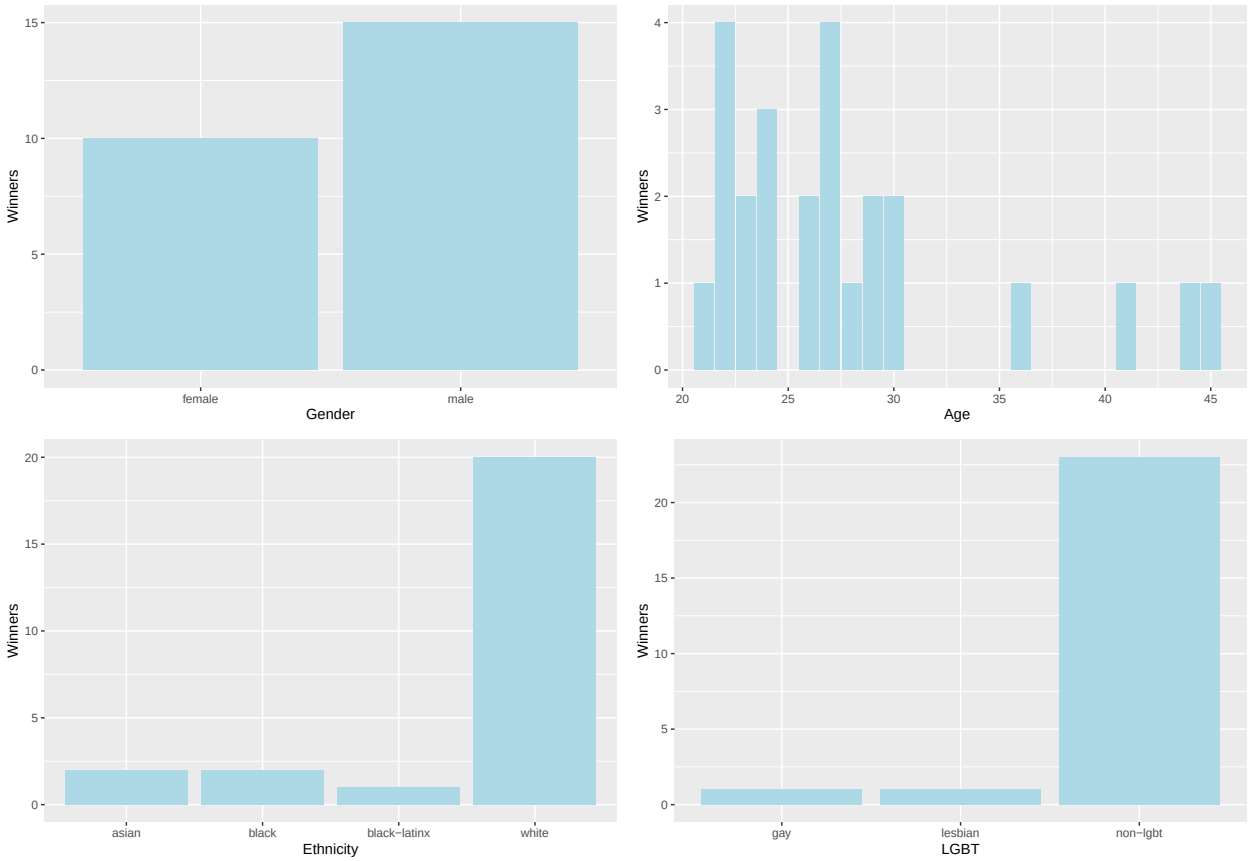
It is interesting to note that my prediction was correct, however. While people of color tend to not last as long in the game in general, in seasons where diversity is high, they tend to perform better. There is a negative correlation between the interaction of poc*season_diversity and final game standings (significant at p=0.004). This is very interesting to see and supports my hypothesis that when the show is able to cast houseguests based on their merit and personality rather than trying to fill a quota for a single person of color and a single lgbtq-identifying person, they cast much more well-rounded people who are able to form relationships and succeed in the game.

What Does A Winner Look Like?

Now that we have seen the factors that indicate a player will perform well in the game, we will look at which factors are specific to indicate a winner.

Data on Past Winners

The demographic breakdown of past winners paints a pretty clear picture. They tend to be male, Caucasian, non-lgbt and in their twenties.



Creating A Model

Since winning or not winning is a binary variable, we will be creating a generalized linear model. I will be analyzing the same factors that I did while looking at overall game performance.

```
##
## Call:
## glm(formula = winner ~ gender + age + poc + poc * season_diversity +
##     comps, family = binomial(), data = houseguests_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5281  -0.3361  -0.2415  -0.1854   2.8017
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.460600   1.129172  -3.065  0.00218 **
## gendermale               -0.060305   0.476384  -0.127  0.89927
## age                      -0.006653   0.033187  -0.200  0.84111
## pocTRUE                  -1.793484   1.847847  -0.971  0.33176
## season_diversity         -1.161914   2.268126  -0.512  0.60846
## comps                     0.519149   0.100645   5.158 2.49e-07 ***
## pocTRUE:season_diversity  4.297757   4.748077   0.905  0.36538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 181.58  on 359  degrees of freedom
## Residual deviance: 145.71  on 353  degrees of freedom
## AIC: 159.71
##
## Number of Fisher Scoring iterations: 6
```

Of all the factors we are looking at, none of the demographic pieces of information are significant in determining a winner. This may seem surprising after looking at the breakdown of winners above. My best guess at why none of the indicators are significant is that coming into the house as a white, Caucasian, non-lgbt twenty-something male does not indicate that a player has a better chance of winning the game, simply because there are so many of them. If there are four houseguests fitting that description on a season, at least three of them would necessarily have to lose.

The only significant statistic that correlates with winning the game is competition wins (significant at p=0.000). This, again, makes sense as winning competitions provides a player with an avenue to protect themselves and their allies. At the end of the game, the jury is most likely to vote for a player who boasts the best in-game resume, and competition wins is a huge part of that.

Testing the Model

While there is a limited amount of data, as the show has only been on air for 23 seasons, future seasons present an opportunity to test the model. The next season begins on July 3rd, 2022 and it would be a good exercise for me to use my models to see if I am able to predict performance for the newly announced houseguests prior to the season.

So How Do I Win the Game?

The only conclusive advice I have to give to aspiring houseguests is: win competitions. Winning competitions is significantly correlated with better performance in the game, making it father, and ultimately winning. It provides a player with the ability to form strong bonds and alliances, especially if those competitions are won early on, and *especially* if they win the first HOH. The other key indicator of performing well on a season is being cast with others that are similar to you. Meaning if you are a person of color, hope that there are others surrounding you. And for good measure, even though there is no statistical correlation, it is probably wise to avoid entering first (only joking, but I am still a bit superstitious).

These models that I have created to predict future houseguests' performance are very much rough drafts that will have to be tested and fine-tuned during future seasons, especially as we lean more and more into diverse and well-rounded casting (not only when it comes to race, but also when it comes to sexuality, gender-norms, and leaning away from TV archetypes).

RMD File

https://github.com/alisssaa/dsc520/blob/master/FinalProject/BB_FinalProject.Rmd

Bibliography

"Big Brother's Diversity Problem." 2019. https://vincedixonportfolio.com/app/big-brother-diversity/.