Final Project Milestone 1

Alissa Trujillo

2022-05-07

Introduction

For my final project in DSC 520, I will be delving into data surrounding the reality show Big Brother. The focus of my research will be to determine which factors statistically influence a house guest's probability of winning. I intend to use the methods learned in this course to analyze what makes a good winner, examining which factors are correlated with making it far in the game. Big Brother has a very involved fan base, it seems that those who watch the show tend to be very passionate about it rather than being casual fans. I think this data would be particularly interesting for those who watch this show or are interested in applying to be on the show. It is a data science problem because I will be analyzing the wealth of data that exists regarding the show and past house guests. I will be taking observations of many variables into account in order to determine trends and correlations that exist in the data, and I will use that to make predictions regarding future house guests and how they will fare in the game.

Research Questions

- Does winning more competitions significantly increase the likelihood of winning the game?
- Winning the first HOH is considered to be a good sign of somebody making it far in the game. Is this true according to the data?
- Are there any predictors or common factors between players who go on to win America's Favorite Player?
- Are there any correlations between players surviving the block (being nominated for eviction but surviving) and lasting longer in the game?
- Do players from a certain locale (certain states, or perhaps big city vs. small city) tend to make it farther in the game?
- How do minority players tend to fare in the game and in competitions?

Approach

I will approach my problem of determining the factors that make a successful Big Brother player by looking at a number of different data sets. I intend to collect more data using the resources that exist online in order to add more variables to the main data set that I found. Using the variables provided as well as others that I am able to discover during my own research, I intend to look at a number of correlations to see what factors highly correlate with winning the game, making it to jury (the final phase of the game), and winning America's Favorite Player. I will also cross-reference to see if these factors are correlated with each other, or if there are significant differences between what makes a player likely to make it far versus winning the whole game. Once I am able to calculate the correlations, I will see if there is any way to prove causation in one direction or another.

Addressing The Problem

This approach will address the problem at hand by providing statistical evidence for whether the beliefs that the audience hold about who is likely to be a winner truly hold up. It is a common belief that winning the first HOH, not walking in the house first, and many other superstitious things are very influential in a house guest ultimately winning the game. I intend on using statistical evidence to prove whether these beliefs are true or are just fallacies that have arisen over the 20 years the show has been on the air.

There is no way to fully address the question on how to, without fault, become a winner of Big Brother. There are many other influences, such as social interactions and random chance introduced by the game itself, that cannot be measured. There is no foolproof way to ensure you will win the game, but I intend to find the most predictive circumstances for which house guests have won and could win the game given the data collected.

Data Sets

Big Brother Contestant Data

https://github.com/vdixon3/big-brother-diversity-data/blob/master/big_brother_data.csv

This will be the main data set I will be using for this project. It includes data on contestant's placing in the game, the number of competitions they have won, whether they made jury or not, where they are from, and many other game-related pieces of information. This data set has 49 variables with 370 observations. This means that there are 49 pieces of information about each of the 370 houseguests that the show has had over the last 23 seasons. The original data was compiled by a github user (vdixon3) in order to do a research project on diversity in the show.

This data has been continuously updated so that it includes the most recent season that concluded in 2021. There are a couple sections of n/a's in the data. The first season did not have HOH or Veto competitions, so that data is n/a as they did not exist. The second season did have HOH competitions but no Veto competitions, so the Veto column is n/a for season 2 as well. America's Favorite Player did not exist until season 7, so all earlier contestants have n/a in that column as well.

Big Brother Diversity Data

https://github.com/vdixon3/big-brother-diversity-data/blob/master/big_brother_data_diversity.csv

This data set was compiled by the same user (vdixon3) as the first data set I have presented. This set of data focuses on the racial diversity of each individual season. It provides breakdowns of each race's representation in the season as well as ratios, biracial inclusion, and separate columns that separate contestants who racial background is unknown. This data set is complete and does not have any n/a's, as it instead puts all house guests with unknown backgrounds into the "unknown" category.

Additional Contestant Data

https://github.com/alisssaa/dsc520/blob/master/FinalProject/data/bb_additional_data.csv

This data set is made up of data I compiled from online sources. I took the first few columns of the contestant data, including their names and identifier codes, so that this data will fit nicely with the data I have already found. I added two additional columns that indicate whether a houseguest was the first to enter the house, and whether they won the first HOH competition. This data was obtained on wikipedia and the Big Brother Wiki. These are two important pieces of information I will need in order to test the hypotheses that those are important indicators for a contestant to win the game. There is one section of n/a values which correspond with the fact that there were no HOH competitions in the first season.

Sources for the data:

first_to_enter
https://bigbrother.fandom.com/wiki/First_to_Enter_Curse

first_hoh
https://bigbrother.fandom.com/wiki/List_of_Head_of_Household_Competitions

Required Packages

Packages that I will need to conduct this project are as follows:

- **ggplot2** in order to make professional looking plots of my data
- **lm.beta** so I will be able to do regression analysis on my variables
- **dplyr** to manipulate data tables as well as combine data from one table with another
- **reshape2** to clean up data tables
- **car** to calculate Cook's distance as well as VIF, perform Durbin-Watson tests and outlier tests, and complete ANOVA tables
- **stats** in order to compute confidence intervals, assess correlation, and fit regression models

Plots and Tables

Since the data is so large, I will use smaller tables to filter the data by season, placement, and other factors to concisely present data that is important to the questions I am presenting to the audience.

I will use scatter plots in order to compare the 370 house guests with each other on various variables. I will overlay them with regression models in order to see if there are trends in data and how far each house guest falls from what the model predicts. Histograms will be useful in seeing how often a certain characteristic (such as winning specific competitions, being part of a specific population, etc.) results in a certain result (such as winning the game or making it to jury).

The tables provided by casewise diagnostics will be particularly useful as well. Residuals vs. Fitted graphs will allow me to see how good the model I develop is based on past data. A Q-Q plot will allow me to see if my data and the residuals are normally distributed. Residuals vs. Leverage graphs will allow me to identify outliers in my data and ensure that the Cook's distance of my data points falls within an acceptable range.

Future Steps

In order to answer my research questions, there is still data I need to collect. I have started collecting data from web sources that will assist me in answering my questions, but I am sure there will be missing pieces that I need to collect later. I also need to work on having a more robust grasp on multiple regression and how to analyze it, as well as having a deeper understanding of what each of the case wise analysis graphs represent and how to interpret them. The housing assignment provided me with a better understanding of them, but I will need more practice in order to fully grasp the shape and function of the graphs and data points. Converting the graphical and numerical information that is produced by the data into descriptions that are useful to an audience is something that I have begun to get more comfortable with, but still need to continue to work at.