Milestone 2

Alissa Trujillo

2022-05-17

Final Project Milestone 2

Importing & Cleaning the Data

First I will import the 3 data sets that I have collected for this project.

```
setwd("/Users/alissa/Documents/Grad/DSC 520/dsc520")

contestant_data <- read.csv("data/bigbrother/contestant_data.csv",
    header = TRUE)
additional_data <- read.csv("data/bigbrother/bb_additional_data.csv",
    header = TRUE)
diversity_data <- read.csv("data/bigbrother/bb_diversity.csv",
    header = TRUE)
```

Two of the data sets, contestant_data and additional_data contain observations regarding each individual houseguest that has participated in the show over the years. These two data sets I aim to combine, taking the most important pieces of each in order to have a concise data set with information that is pertinent to answering my research question: what factors are highly correlated with success in the Big Brother game?

The third data set, diversity_data, contains important demographic information about each of the show's seasons. This data set is separated by season rather than individual player like the other two. I will therefore need to take the season data and apply it to each of the season's houseguests so I am able to use it in my final dataset. My first step in cleaning my data is taking some diversity information and transforming it so that it gives each houseguest a variable for their season's diversity. I did this using excel as it was more intuitive to me. The values are shown below, originally in my diversity data set and secondly in my additional data dataset.

```
diversity_data[, c("season_code", "poc_percent")]
```

```
##    season_code poc_percent
## 1        bbus1        0.30
## 2        bbus2        0.17
## 3        bbus3        0.17
## 4        bbus4        0.31
## 5        bbus5        0.07
## 6        bbus6        0.29
## 7        bbus7        0.21
## 8        bbus8        0.14
## 9        bbus9        0.13
## 10      bbus10        0.31
```

```
## 11      bbus11        0.38
## 12      bbus12        0.15
## 13      bbus13        0.36
## 14      bbus14        0.13
## 15      bbus15        0.19
## 16      bbus16        0.31
## 17      bbus17        0.18
## 18      bbus18        0.56
## 19       bbott        0.23
## 20      bbus19        0.35
## 21      cbbus1        0.36
## 22      bbus20        0.38
## 23      cbbus2        0.50
## 24      bbus21        0.31
## 25      bbus22        0.31
## 26      bbus23        0.50
```

```
head(additional_data[, c("first", "last", "season_code", "season_diversity")])
```

```
##       first    last season_code season_diversity
## 1  Brittany  Petros       bbus1              0.3
## 2 Cassandra  Waldon       bbus1              0.3
## 3    Curtis     Kin       bbus1              0.3
## 4     Eddie   McGee       bbus1              0.3
## 5    George Boswell       bbus1              0.3
## 6     Jamie    Kern       bbus1              0.3
```

I intend to use this information to see if competing in a diverse season has an effect on how well players of color do in the game. Does it make them more likely to make jury or even win the game?

Now that I have extracted the information I need from the diversity data set, I am going to combine the contestant_data and additional_data data frames. The contestant_data dataset includes 49 variables, some of which are not important to answering my research question. The additional_data dataset includes 7 variables, 3 of which are identical to contestant_data. I am only going to take the most important pieces of information in order to create a new, concise data frame. Because I originally created the base of additional_data using the first 3 columns (first, last, season code) of contestant_data, I know that they match up correctly and I should have no problems combining the information.

```
houseguests_df <- subset(contestant_data, select = c("first",
    "last", "season_code"))
houseguests_df$age <- contestant_data$age
houseguests_df$gender <- contestant_data$gender
houseguests_df$ethnicity <- contestant_data$race_ethnicity
houseguests_df$poc <- contestant_data$race_ethnicity != "white"
houseguests_df$season_diversity <- additional_data$season_diversity
houseguests_df$appearance <- contestant_data$appearance
houseguests_df$first_in_house <- as.logical(additional_data$first_in_house)
houseguests_df$comps <- contestant_data$total_wins
houseguests_df$hoh <- contestant_data$total_hoh
houseguests_df$veto <- contestant_data$total_vetos
houseguests_df$first_hoh <- additional_data$first_hoh == 1
houseguests_df$nom <- contestant_data$total_nominations
houseguests_df$afp <- contestant_data$afp
```

```
houseguests_df$jury <- contestant_data$made_jury
houseguests_df$jurymember <- houseguests_df$jury == "yes" |
    houseguests_df$jury == "final"
houseguests_df$placement <- contestant_data$final_placement
houseguests_df$final2 <- houseguests_df$jury == "final"
houseguests_df$runnerup <- contestant_data$final_placement ==
    2
houseguests_df$winner <- contestant_data$final_placement ==
    1
```

Here I begin by creating a new data frame that contains the 3 columns that the two data sets have in common. Then I have added the additional variables that I would like to examine that I think are relevant to a contestant's success. The first few variables are just demographic information. The appearance variable denotes if this is a houseguest's first, second, etc. time competing on the show. I intend to see if multiple appearances give a player an advantage in the game. The first_in_house variable indicates whether a houseguest was the first to enter the house at the beginning of the season. There is a widely held belief that it is a "curse" to walk in first, and I will be using it to see if there is a statistically significant hindrance to a contestant's final placing.

The next few variables (comps, hoh, veto) will tell us how many and which kind of competitions each player wins. Winning competitions gives you power in the game, but when players amass too much power, they increase their threat-level and can end up getting targeted by the other contestants. I will hypothesize that winning competitions will be correlated with making it far in the game, but not necessarily winning. The first_hoh variable denotes whether a player won the very first HOH of the season. This is widely believed by fans to be an indicator of a winner, as it is an opportunity to make connections the first week, giving an advantage moving forward. I want to test whether this belief is statistically backed or not.

The final few variables are indicators of a houseguest's success in the game. The jury variable tells us whether the houseguest made jury, meaning they made it to the halfway point and had a vote in determining the winner of the game. The afp variable tells us whether the player won America's Favorite Player, meaning that they were the most universally-liked houseguest of the season by the audience. This designation does not begin until season 7, so the data is more limited. The placement variable tells us the houseguest's final placement in the game, and lastly the winner variable denotes if they won the season.

The final data set includes 22 variables, which is much more manageable than the copious amount of information that we started with in the original three data sets.

The last transformation I am going to do is to remove the season 1 data from the table. The first season was filmed in a completely different format, where viewers voted on their favorite and least favorite players based on entertainment value. There were no competitions, no type of power available, and no necessity for social relationships. Because of the nature of the season, I do not think the data is relevant to the research I am conducting.

```
houseguests_df <- subset(houseguests_df, season_code != "bbus1")
```

The Final Dataset

```
head(houseguests_df)
```

```
##     first       last season_code age gender ethnicity   poc season_diversity
## 11   Will      Kirby       bbus2  28   male     white FALSE             0.17
## 12 Nicole Schaffrich       bbus2  31 female     white FALSE             0.17
## 13 Monica     Bailey       bbus2  40 female     black  TRUE             0.17
```

```
## 14  Hardy  Ames-Hill       bbus2  31   male     white FALSE             0.17
## 15  Bunky     Miller       bbus2  36   male     white FALSE             0.17
## 16 Krista    Stegall       bbus2  28 female     white FALSE             0.17
##    appearance first_in_house comps hoh veto first_hoh nom afp  jury jurymember
## 11          1          FALSE     0   0   na     FALSE   4  na final       TRUE
## 12          1           TRUE     1   1   na     FALSE   2  na final       TRUE
## 13          1          FALSE     1   1   na     FALSE   2  na   yes       TRUE
## 14          1          FALSE     3   3   na     FALSE   1  na   yes       TRUE
## 15          1          FALSE     0   0   na     FALSE   1  na   yes       TRUE
## 16          1          FALSE     1   1   na     FALSE   2  na   yes       TRUE
##    placement final2 runnerup winner
## 11         1   TRUE    FALSE   TRUE
## 12         2   TRUE     TRUE  FALSE
## 13         3  FALSE    FALSE  FALSE
## 14         4  FALSE    FALSE  FALSE
## 15         5  FALSE    FALSE  FALSE
## 16         6  FALSE    FALSE  FALSE
```

Learning Opportunities

I was able to easily take the data from the diversity data set, which was separated by season rather than houseguest, and append it to each houseguest from the season using Excel. However, I was a bit overwhelmed when initially attempting to do it via R. Taking a value and applying it to every data point that possesses a certain value second variable is definitely something that is doable using R, but I wasn't able to figure it out using the documentation. There are definitely a number of small things like this that I do not find intuitive to do using R that I need to get more comfortable with. I am thankful that I have the tools and experience to be able to manipulate data in other ways but I would like to become more proficient doing so using R.

Information That Is Not Self-Evident

I believe most of the variables are pretty self-explanatory, but it will be interesting to see what kind of interactions exist between the variables. I know I want to look at the interaction between being a person of color and being on a season that has high vs. low racial diversity, as I think that will give a more rich understanding than looking at the variables on their own. The variable season_diversity is pretty useless on its own, as it is the same for everyone in the season regardless of placement.

Looking At the Data

The ultimate question I am researching, is what factors are correlated with being a successful Big Brother player? While this, of course, includes winning the game, there are a couple other accomplishments that are considered an indicator of success in the game. Making it to the jury phase of the game shows that a player was able to make important social connections in order to get to the halfway point of the game. I am interested in seeing if there are commonalities or differences in what factors indicate making it far in the game versus winning the whole season.

I would also like to look at data surrounding runner up players. The people who end up in second place typically are well-liked but considered to be non-threatening, they are typically brought to the end by stronger players who think they can beat them. I am curious to see if there are any trends in predicting second-place players, perhaps lack of competition wins. The last indicator of success that is measured is America's Favorite Player, I am curious to see if any factors correlate with America's perception of a player. I have a hunch that the qualities that lead to likability are tied to personality rather than the variables measured here, but it will be interesting to look into regardless.

Slicing & Dicing Data

I have created a number of new variables while putting together this final data set. I created a true/false variable denoting whether a player is a person of color by analyzing the text in the ethnicity variable. I also created boolean variables indicating whether a player won the game or was a runner-up. I converted the first_in_house and first_hoh variables to true/false rather than the binary variables they were originally coded as.

I combined two of the three original datasets that I began with in the first milestone, contestant_data and additional_data. I took pieces of information from the third data set, bb_diversity, to include diversity and ethnicity information in my main data table.

I also removed the season 1 data from houseguest_df as it is missing many crucial pieces of information (due to the show being filmed in a very different way) and does not provide information relevant to the questions I am asking.

When I begin to analyze success factors, I will have to take into account that certain things did not come into play until later seasons. The veto competition did not take place until season 3 and America's favorite player was not voted on until season 7, so I will have to take subsets of the data or find other ways to deal with the na's in the data when I am analyzing these variables.

Summarizing Data

I intend on performing a number of regressions to create models to test my hypotheses. I will first do a linear regression in order to see what variables are correlated with overall placement on the show, which is a numerical variable. Then I intend to make logistical regression models to assess the binary success variables that I have (winning or not winning, making jury or not making jury, winning America's favorite player or not). I will then compare and contrast to see if there are any indicators that are uniform across all measures of success, or if they are all predicted by different characteristics.

Plots and Tables

I plan to first use basic histograms and scatter plots to start off looking at the interaction of some of the variables in the data, for example, to see how diversity has changed over the course of the show or to see the discrepancy between female and male winners. I think it is important for the audience to get an understanding of the show and its demographics before I get into the nitty gritty statistical stuff.

Then once I get deeper into the model creation, I will use the casewise diagnostic plots in order to assess my models and assumptions about my data. I will use Q-Q plots to see if my data is normally distributed and Residuals vs. Fitted graphs to see if my model is unbiased.

I intend to use tables in order to display important information to the reader, including snippets of data and model summaries.

Learning Opportunities

The most important thing that I need to learn to be able to successfully answer my question is how to prove causation rather than just correlation. Being able to establish a correlation between variables to support my hypothesis will be sufficient, but it would be even more satisfying to be able to actually prove directional causation. In this particular case, the winner is announced at the culmination of the game, so temporally all of the other variables had to come first. But I am not sure if that is enough to truly establish causation. That also does not hold true for my jury variable, since jury is the midpoint of the game.

I have become much more comfortable with regression models over the last few weeks, but I still find myself looking for guidance when analyzing results and drawing conclusions from the numbers and graphs. Also,

I need to learn to construct tables in order to showcase the data. There were a couple examples in the textbook but I am unsure if it is a command in R or if they are just created by hand.

Machine Learning

I am not entirely confident about what machine learning really encapsulates, but I hope to be able to use the tools I learn in the last few weeks in the class to help myself create models in an efficient way. Rather than trying numerous combinations of variables in order to find the best model, I would love to be able to use machine learning in order for the model to determine those variables.