Survivor Data Project

What Makes the Best Contestant

Annie Walsh, Audrey Sauter, Shane Matsushima
Alex J. Quijano
MTH 361: Applied Statistics
27 April 2023
University of Portland

Table of Context

Introduction

Outwit, Outlast, Outplay and become Sole Survivor. For 23 years and 43 seasons later, individuals compete in physical and mental challenges in a complex social game on the remote islands of Fuji. Players of all different ages from diverse backgrounds play for the chance to win 1 million dollars, but what makes one person out of 18 castaways the Sole Survivor? In this study, we are interested in the statistics to discover what makes the best contestant on Survivor. We looked at four main aspects of contestants to determine if the rate of challenge wins, being a returning player, age, and personality type are the key components to winning Survivor. We used histograms, scatter plots, means tests, and z-scores to analyze our data and see relations between these characteristics and the winners of Survivor. We hypothesized that certain characteristics and traits (i.e., age, previous wins, relationships with other contestants) of contestants might correlate with their likelihood of winning the competition.

Data Exploration

The Data Set

There are 17 RDA files in the survivoR dataset. Each has different dimensions and different variables. One example data frame we are using is the castaway's data frame. In this data frame, there are 16 variables and 1185 observations. The data comes from all seasons and over 40 participants in the show. The sampling strategy we will use is Stratified Sampling, as each RDA file could be recognized as a different subpopulation for the show. The following table is an example of the castaway variable descriptions:

version	The version of the show that the castaway was participating in. This would be between AU, US, SA, and NZ [Character]
version_Season	ID for which version and episode the castaway data is being taken from [Character]
season_name	Name of the season that the castaway was a part of. [Character]
season	Season number of observation [Numeric]
full_name	Full name of the castaway [Character]
castaway_id	Specific ID for the castaway based on the version [Character]

castaway	First name or nickname of the castaway [Character]
age	Age of the castaway [Numeric]
city	Origin city that the castaway is from [Character]
state	State the castaway is from [Character]
episode	Episode the castaway was sent home [Numeric]
Day	Number of days the castaway stayed on the show? [Numeric]
Order	Order in which the castaway was voted out [Numeric]
result	Reasoning for when or why the castaway was sent home [Character]
jury_status	Describes which number of the jury the castaway was apart of (NA indicates they were not selected to be in the jury) [Character]
original_tribe	The tribe the castaway started off with [Character]

Exploration

Prior to diving into the hypothesis testing of the data set, in order to produce a valuable hypothesis, the team explored the data set by plotting and calculating different attributes in order to get a better understanding of what the data was being used for.

The first scatter plot and table (Appendix A1 and A2) are the outcomes of exploring age between castaways on the show. Some aspects of the data that was seen is that the minimum age of a contestant was 18, with an average age of 33, but the max age (or oldest age on the show) being 75. Through the scatterplot (Appendix A1) the majority of contestants do lie between the ages of 22 to 35.

In Appendix B, the bar graph indicates the number of challenges based on its category. The majority of challenges being used on the show are race challenges, with puzzle, balance, and precision challenges being more commonly used as well. The graph does show a large deficit between race challenges and the rest of the challenges, meaning that is very likely to encounter more race challenges then any other challenge category.

Appendix C showcases the number of challenges won per castaway. Appendix C1 showcases the spread of how many contestants won a specific number of challenges. With the help of the table in Appendix C2, the data found was that the minimum of challenges won was 0, with a max of 42. The average amount of challenges won was around 7 to 8. This is also indicated in the bar graph (Appendix C1) as there are higher counts of castaways that won 5 to 9 challenges, with another large spike in castaways only winning close to 2 to 3. The bar graph is shown as being right skewed, with there being little to no participants winning more than 30 games.

The final exploration the team looked at was the total number of challenges per castaway based on a number of the challenge wins per castaway. The scatter plot graph with linear regression line (Appendix D) showcases an upward trend of total challenges and the number of those challenges being won. As the number of challenges increase, the amount of wins also increases. The slope of the line was 0.408, meaning on average, castaways win 40.8% of the challenges they compete in.

Methods

Hypothesis

By analyzing data from previous seasons of Survivor winners, we hypothesize that certain characteristics and traits (i.e. age, previous wins, relationships with other contestants) of contestants may correlate with their likelihood of winning the competition. **Does the rate of challenge wins affect your chances of winning? If you are a returning player, are you more likely to be voted off? Does your age impact your chances of winning?** By identifying these correlations, we aim to develop a statistical model that can predict the success of potential contestants and suggest the optimal combination of characteristics for the ideal Survivor contestant

In order to find what makes the best contestant for the show, the team created four hypothesis tests based on the research questions found from the exploration.

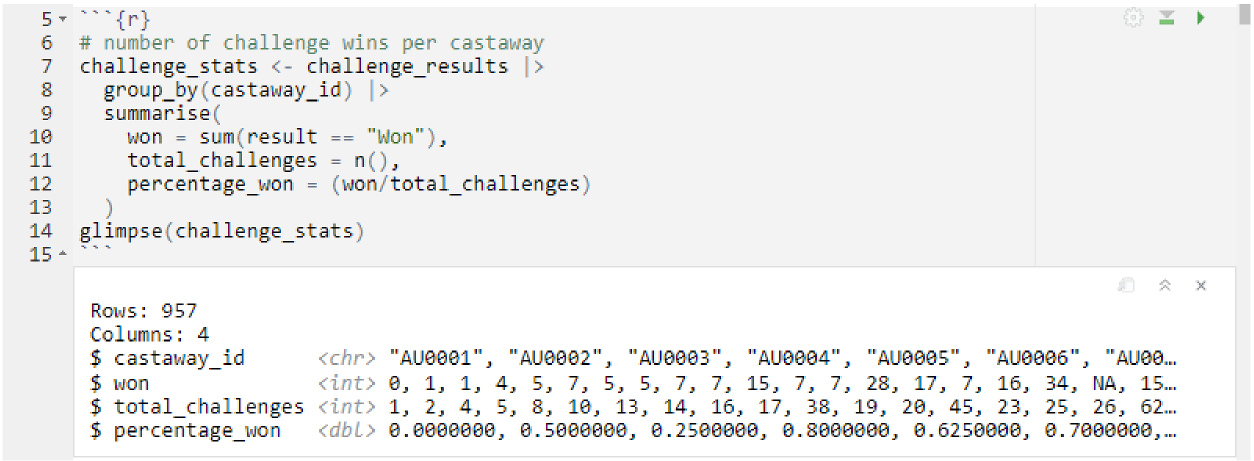
- Does the rate of challenge wins affect your chances of winning?
    a. *Null_Hypothesis*: The rate of challenge wins has no effect on your chances of winning.
    b. *Alternative_Hypotehsis*: the rate of challenge wins has an effect on your chances of winning.
- If you are a returning player, are you more likely to be voted off?
    a. *Null_Hypothesis*: Being a returning player will not affect winning the game.
    b. *Alternative_Hypotehsis:* Being a returning player has an effect on your chances of winning.
- Does your age impact your chances of winning?
    a. *Null_Hypothesis:* Age does not have an impact on chances of winning
    b. *Alternative_Hypothesis:* Age does have an impact on chances of winning
- Does personality type affect the chances of winning the show
    a. *Null_Hypothesis*: Personality type does not impact chances of winning
    b. *Alternative_Hypothesis*: Personality type does have an impact on chances of winning

Hypothesis Testing

Does the rate of challenge wins affect your chances of winning?

What is the average percentage of challenge wins per castaway?

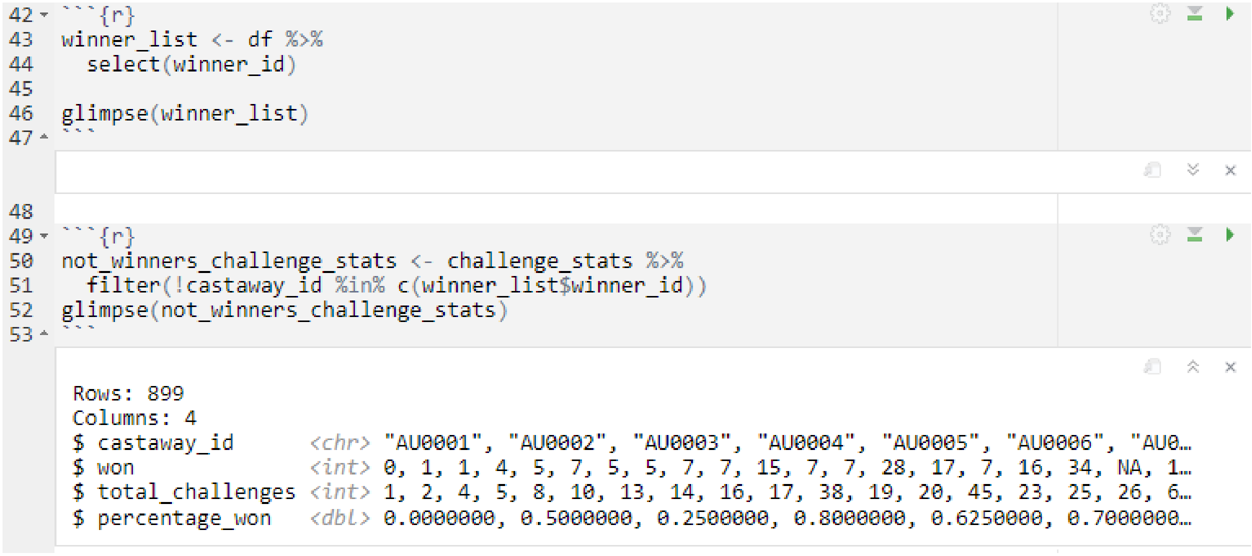
The first step was to create a table that had the following statistics for each castaway: number of challenge wins, number of challenges competed in, and the percentages of challenges won (wins/total). Then filter out the data for contestants who won their season.



~~~
```{r}
# number of challenge wins per castaway
challenge_stats <- challenge_results |>
  group_by(castaway_id) |>
  summarise(
    won = sum(result == "Won"),
    total_challenges = n(),
    percentage_won = (won/total_challenges)
  )
glimpse(challenge_stats)
```
~~~

```
Rows: 957
Columns: 4
$ castaway_id      <chr> "AU0001", "AU0002", "AU0003", "AU0004", "AU0005", "AU0006", "AU00…
$ won              <int> 0, 1, 1, 4, 5, 7, 5, 5, 7, 7, 15, 7, 7, 28, 17, 7, 16, 34, NA, 15…
$ total_challenges <int> 1, 2, 4, 5, 8, 10, 13, 14, 16, 17, 38, 19, 20, 45, 23, 25, 26, 62…
$ percentage_won   <dbl> 0.0000000, 0.5000000, 0.2500000, 0.8000000, 0.6250000, 0.7000000,…
```

Then filter out the data for contestants who won their season.

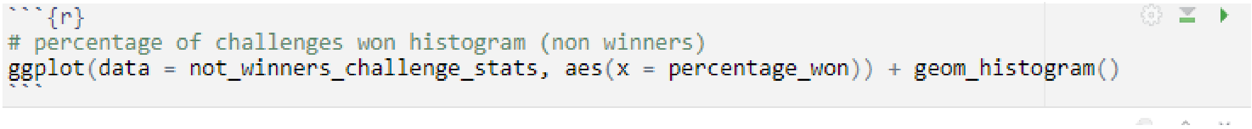


~~~
```{r}
winner_list <- df %>%
  select(winner_id)

glimpse(winner_list)
```
~~~

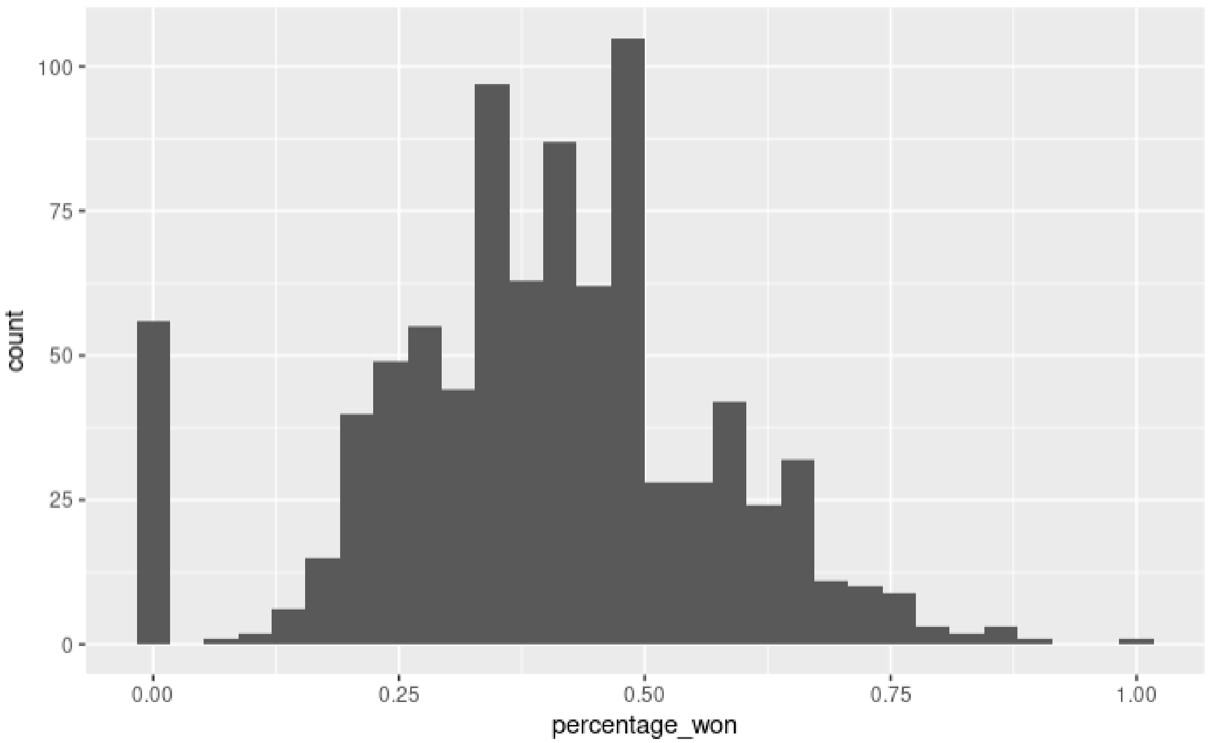
~~~
```{r}
not_winners_challenge_stats <- challenge_stats %>%
  filter(!castaway_id %in% c(winner_list$winner_id))
glimpse(not_winners_challenge_stats)
```
~~~

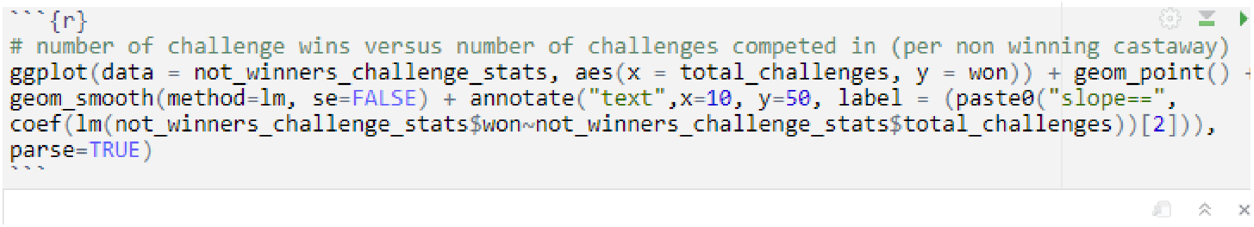
```
Rows: 899
Columns: 4
$ castaway_id      <chr> "AU0001", "AU0002", "AU0003", "AU0004", "AU0005", "AU0006", "AU0…
$ won              <int> 0, 1, 1, 4, 5, 7, 5, 5, 7, 7, 15, 7, 7, 28, 17, 7, 16, 34, NA, 1…
$ total_challenges <int> 1, 2, 4, 5, 8, 10, 13, 14, 16, 17, 38, 19, 20, 45, 23, 25, 26, 6…
$ percentage_won   <dbl> 0.0000000, 0.5000000, 0.2500000, 0.8000000, 0.6250000, 0.7000000…
```

Then create a histogram and scatter plot to represent this table.

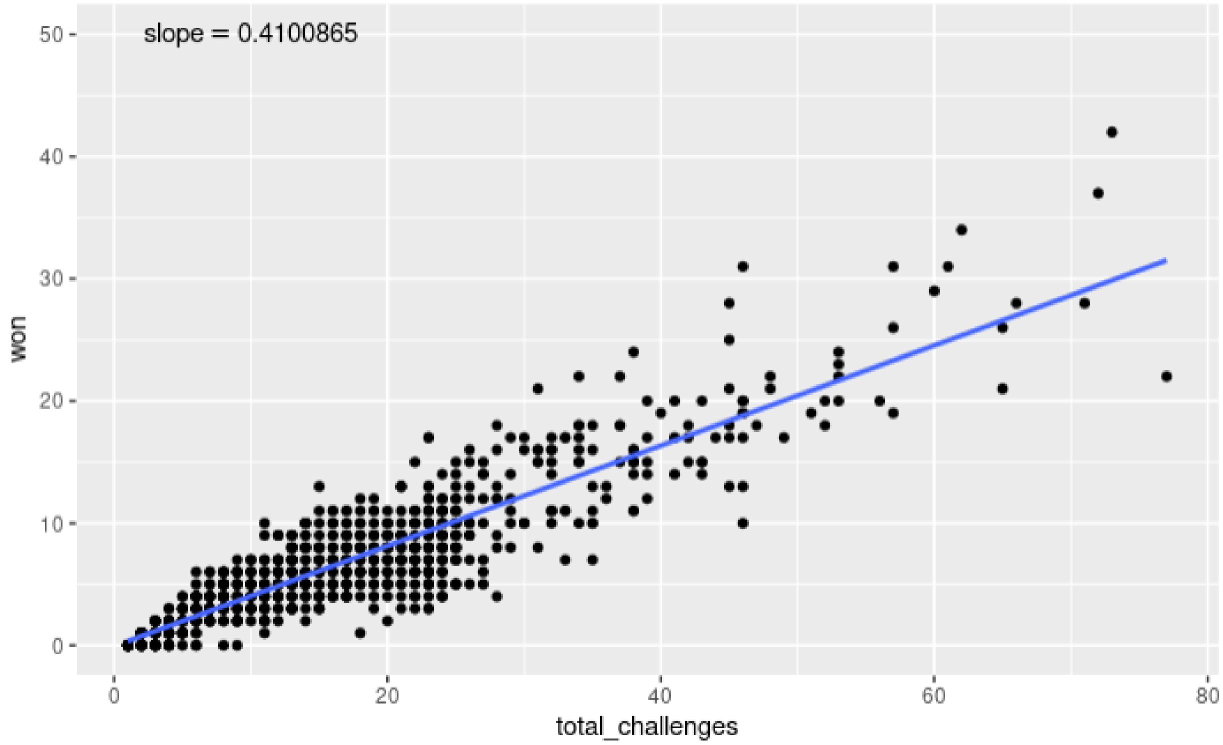


~~~
```{r}
# percentage of challenges won histogram (non winners)
ggplot(data = not_winners_challenge_stats, aes(x = percentage_won)) + geom_histogram()
```
~~~

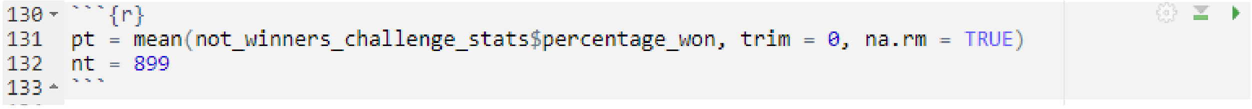




```{r}
# number of challenge wins versus number of challenges competed in (per non winning castaway)
ggplot(data = not_winners_challenge_stats, aes(x = total_challenges, y = won)) + geom_point() +
geom_smooth(method=lm, se=FALSE) + annotate("text",x=10, y=50, label = (paste0("slope==",
coef(lm(not_winners_challenge_stats$won~not_winners_challenge_stats$total_challenges))[2])),
parse=TRUE)
```



To find the average percentage of wins we used the following code:

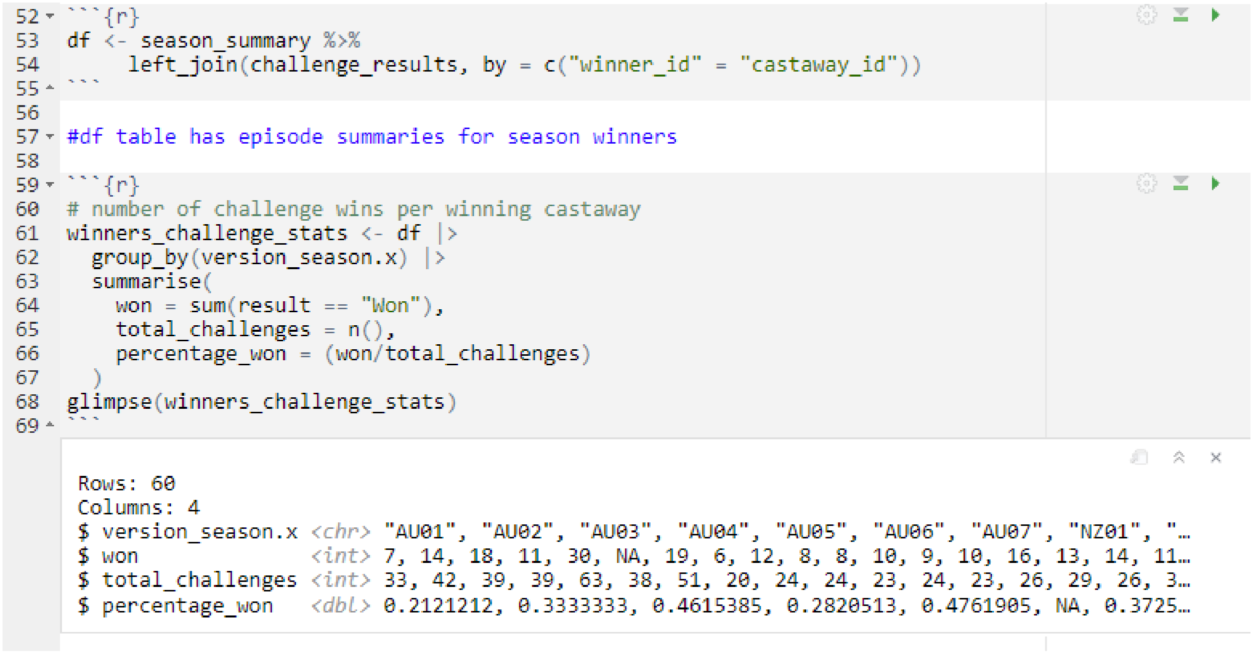


~~~
```{r}
pt = mean(not_winners_challenge_stats$percentage_won, trim = 0, na.rm = TRUE)
nt = 899
```
~~~

The average percentage of challenge wins per survivor is 39.4871%

What is the average percentage of challenge wins per season winner?

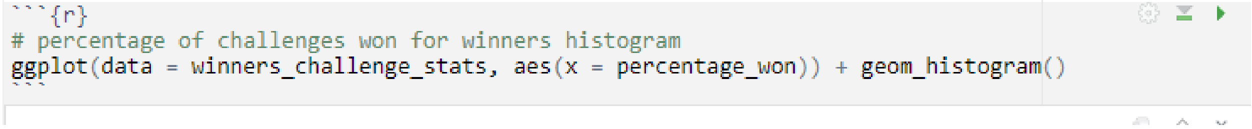
Next was to isolate the season winners to compare their average percentage of challenge wins. The first step was to create a new table to isolate the season winners:



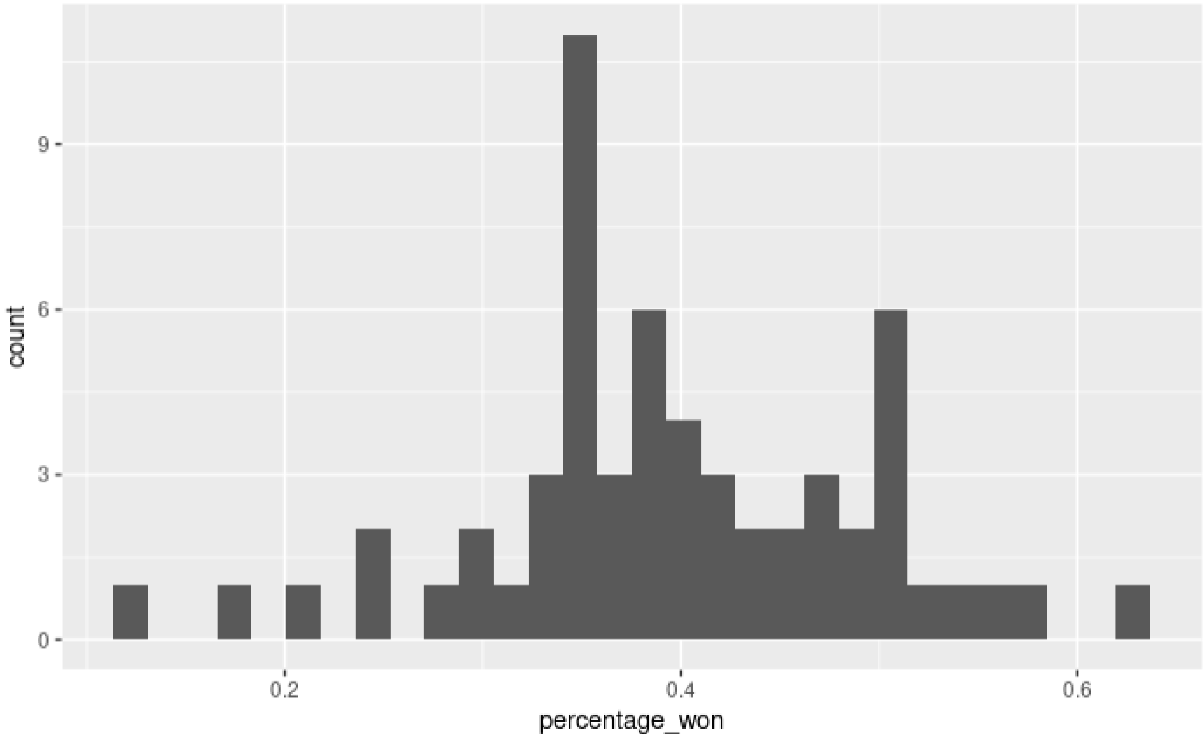
~~~
```{r}
df <- season_summary %>%
     left_join(challenge_results, by = c("winner_id" = "castaway_id"))
```

#df table has episode summaries for season winners

```{r}
# number of challenge wins per winning castaway
winners_challenge_stats <- df |>
  group_by(version_season.x) |>
  summarise(
    won = sum(result == "Won"),
    total_challenges = n(),
    percentage_won = (won/total_challenges)
  )
glimpse(winners_challenge_stats)
```
~~~

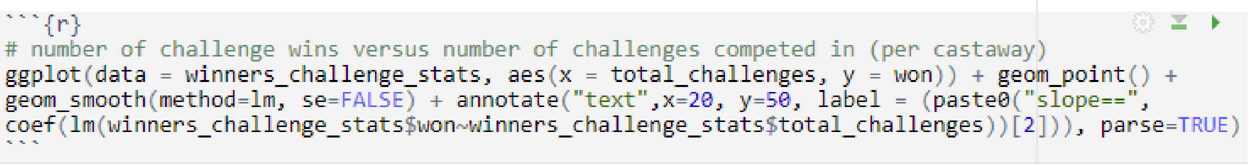
```
Rows: 60
Columns: 4
$ version_season.x <chr> "AU01", "AU02", "AU03", "AU04", "AU05", "AU06", "AU07", "NZ01", "…
$ won              <int> 7, 14, 18, 11, 30, NA, 19, 6, 12, 8, 8, 10, 9, 10, 16, 13, 14, 11…
$ total_challenges <int> 33, 42, 39, 39, 63, 38, 51, 20, 24, 24, 23, 24, 23, 26, 29, 26, 3…
$ percentage_won   <dbl> 0.2121212, 0.3333333, 0.4615385, 0.2820513, 0.4761905, NA, 0.3725…
```

Then create a histogram and scatter plot to represent this table.

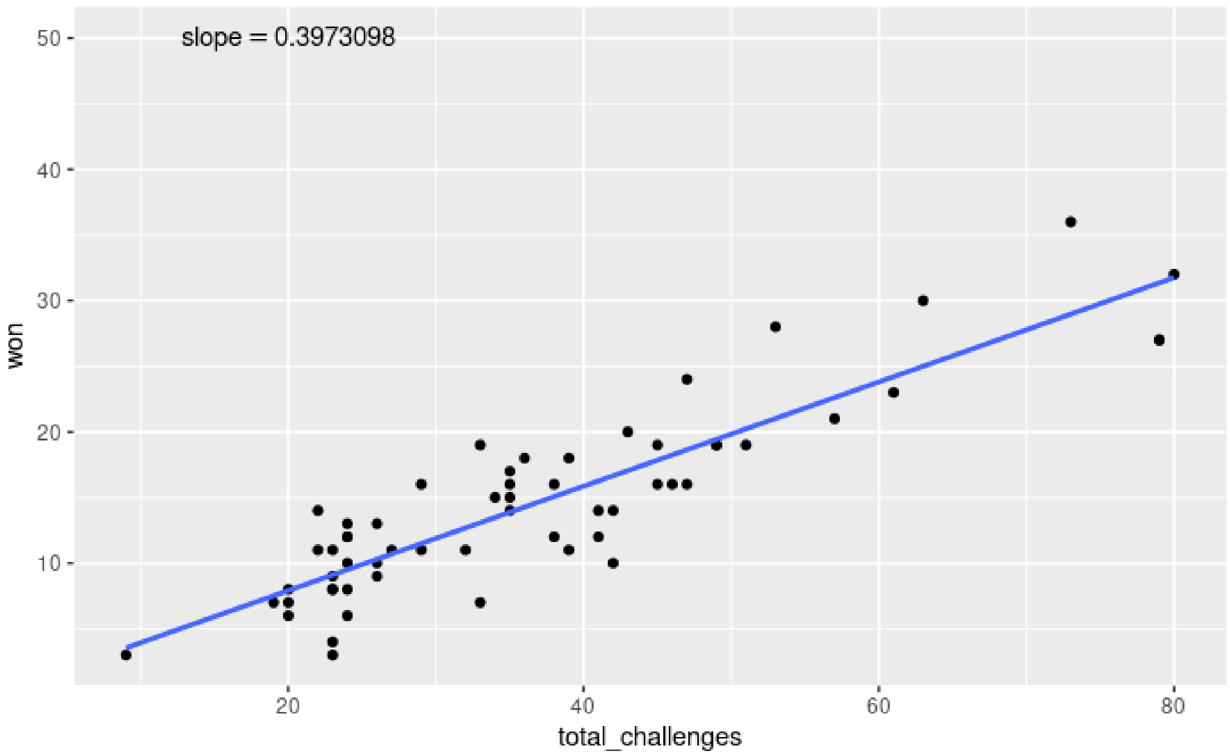


~~~
```{r}
# percentage of challenges won for winners histogram
ggplot(data = winners_challenge_stats, aes(x = percentage_won)) + geom_histogram()
```
~~~

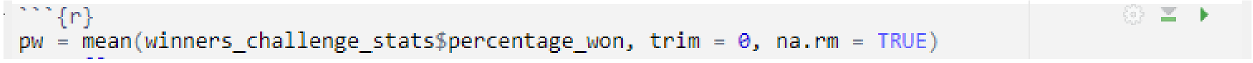




~~~
```{r}
# number of challenge wins versus number of challenges competed in (per castaway)
ggplot(data = winners_challenge_stats, aes(x = total_challenges, y = won)) + geom_point() +
geom_smooth(method=lm, se=FALSE) + annotate("text",x=20, y=50, label = (paste0("slope==",
coef(lm(winners_challenge_stats$won~winners_challenge_stats$total_challenges))[2])), parse=TRUE)
```
~~~



To find the average percentage of wins for season winners, using the following code:



~~~
```{r}
pw = mean(winners_challenge_stats$percentage_won, trim = 0, na.rm = TRUE)
~~~

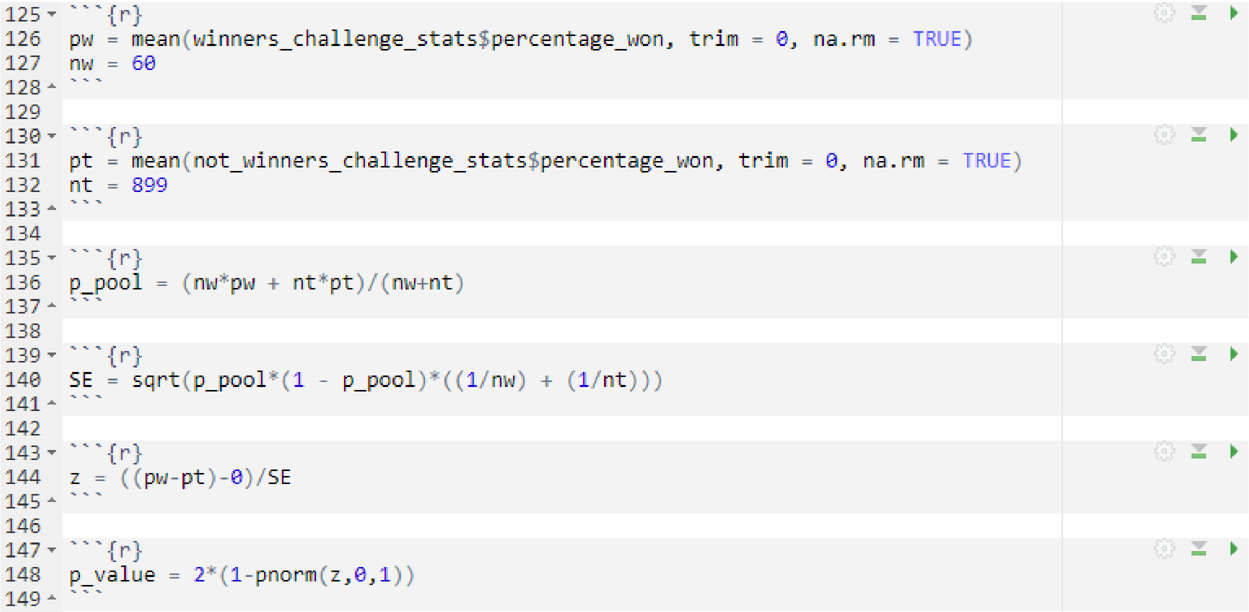
The average percentage of challenge wins per season winner is 39.5174%

To test the hypothesis (Does the rate of challenge wins affect your chances of winning?)The null and alternative hypothesis were formatted as follows:

Null_Hypothesis: The average rate of challenge wins for non-winning contestants is equal to the rate of challenge wins for season winners.

Alternative_Hypotehsis: The average rate of challenge wins for non-winning contestants is not equal to the rate of challenge wins for season winners.

To test this hypothesis, we used an inference for the difference of 2 proportions test (shown below)



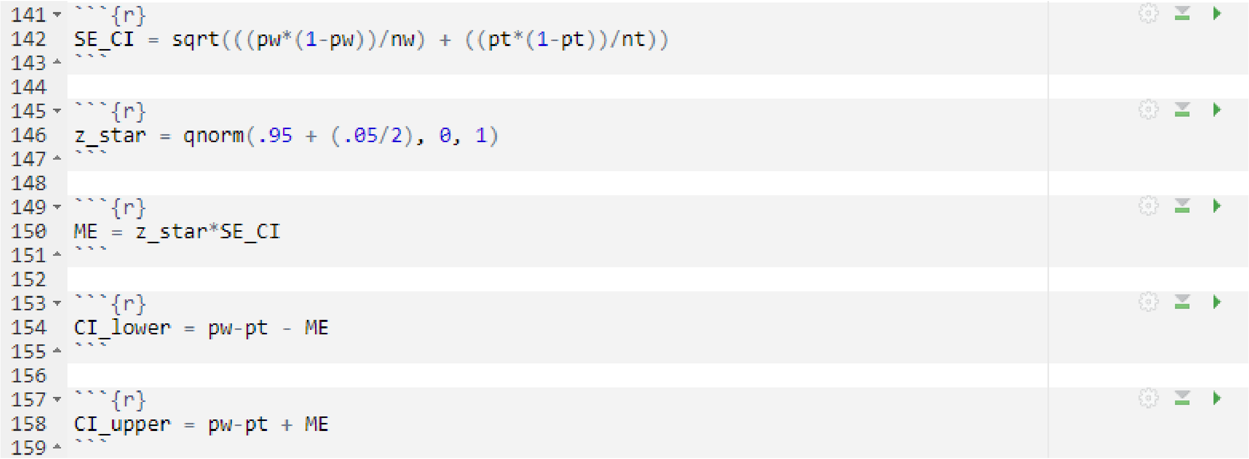
~~~
```{r}
pw = mean(winners_challenge_stats$percentage_won, trim = 0, na.rm = TRUE)
nw = 60
~~~

```{r}
pt = mean(not_winners_challenge_stats$percentage_won, trim = 0, na.rm = TRUE)
nt = 899
```

```{r}
p_pool = (nw*pw + nt*pt)/(nw+nt)
```

```{r}
SE = sqrt(p_pool*(1 - p_pool)*((1/nw) + (1/nt)))
```

```{r}
z = ((pw-pt)-0)/SE
```

```{r}
p_value = 2*(1-pnorm(z,0,1))
```
```

This resulted in a p-value of .996 which means there is not enough evidence to suggest a difference in proportions.

**To then test for error**, a confidence interval, using a confidence level of 95%, was constructed :



```
```{r}
SE_CI = sqrt(((pw*(1-pw))/nw) + ((pt*(1-pt))/nt))
```

```{r}
z_star = qnorm(.95 + (.05/2), 0, 1)
```

```{r}
ME = z_star*SE_CI
```

```{r}
CI_lower = pw-pt - ME
```

```{r}
CI_upper = pw-pt + ME
```
```

The confidence interval is (-.1274, .128)
Because 0 is contained in the interval, there is no significant data to suggest the proportions are different.Thus the null hypothesis was true.
```

If you are a returning player, are you more likely to win the game?

To test the Hypothesis

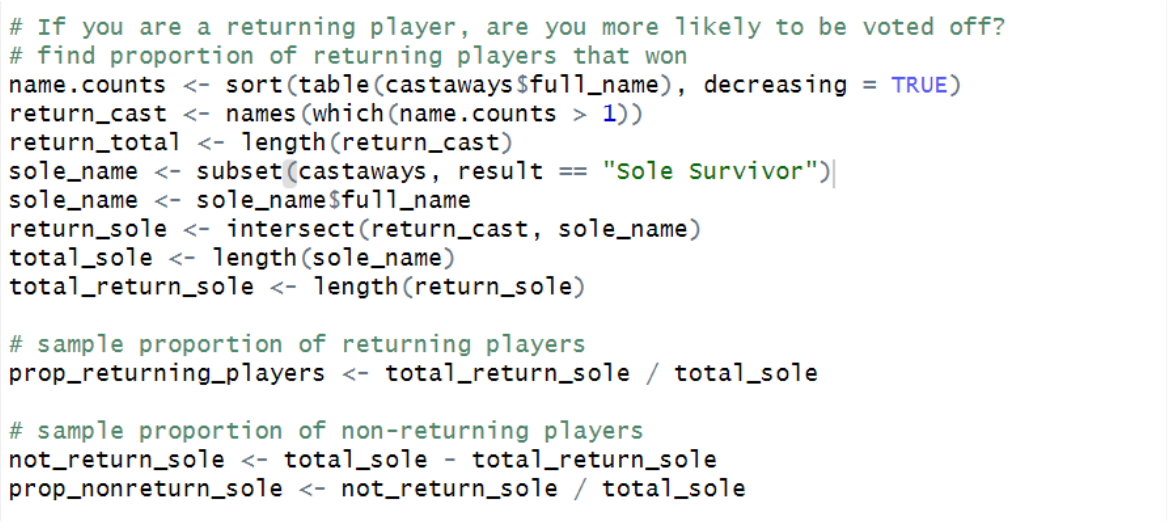
The following are the hypothesis for the test (alpha is 5% or 0.05):

Null_Hypothesis: Being a returning player will not affect winning the game.

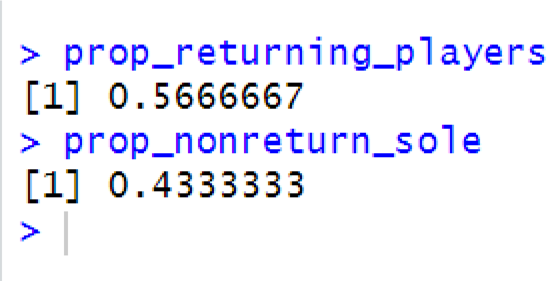
Alternative_Hypotehsis: Being a returning player has an effect on your chances of winning.

Calculating Sample proportion

In order to run a z-score test on the hypothesis, we calculated the sample proportions of non-returning winners and the proportion of returning winners by doing the following:

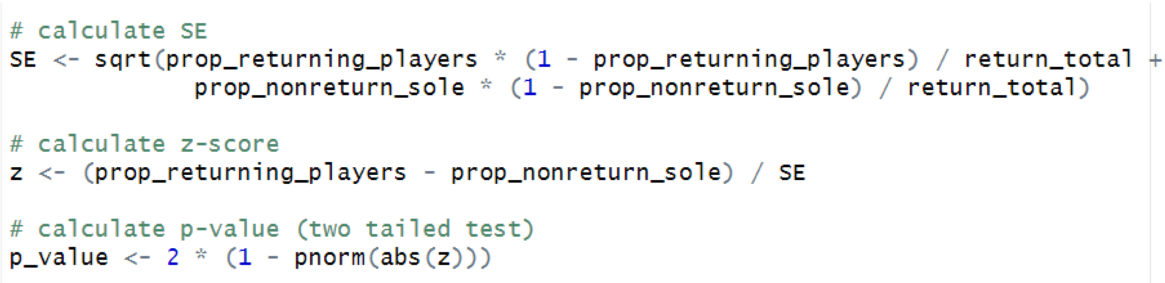


```r
# If you are a returning player, are you more likely to be voted off?
# find proportion of returning players that won
name.counts <- sort(table(castaways$full_name), decreasing = TRUE)
return_cast <- names(which(name.counts > 1))
return_total <- length(return_cast)
sole_name <- subset(castaways, result == "Sole Survivor")
sole_name <- sole_name$full_name
return_sole <- intersect(return_cast, sole_name)
total_sole <- length(sole_name)
total_return_sole <- length(return_sole)

# sample proportion of returning players
prop_returning_players <- total_return_sole / total_sole

# sample proportion of non-returning players
not_return_sole <- total_sole - total_return_sole
prop_nonreturn_sole <- not_return_sole / total_sole
```

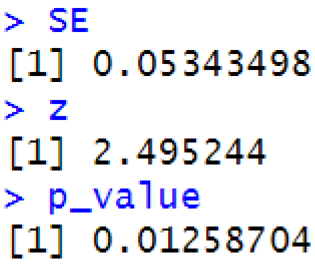


```
> prop_returning_players
[1] 0.5666667
> prop_nonreturn_sole
[1] 0.4333333
> 
```

Once proportions found for prop_returning_players was 0.5666 or 56.66% while prop_nonreturn_sole was 0.4333 or 43.33%. After determining the proportions, we can proceed to generate a z-score and calculate the p-value, along with obtaining the standard error using the subsequent code:



```
# calculate SE
SE <- sqrt(prop_returning_players * (1 - prop_returning_players) / return_total +
             prop_nonreturn_sole * (1 - prop_nonreturn_sole) / return_total)

# calculate z-score
z <- (prop_returning_players - prop_nonreturn_sole) / SE

# calculate p-value (two tailed test)
p_value <- 2 * (1 - pnorm(abs(z)))
```



```
> SE
[1] 0.05343498
> z
[1] 2.495244
> p_value
[1] 0.01258704
```

The z-score found was 2.4952, so the p-value found was 0.012, which is less than the alpha of 5%, or 0.05.

Does your age impact your chances of winning?

To test the hypothesis:

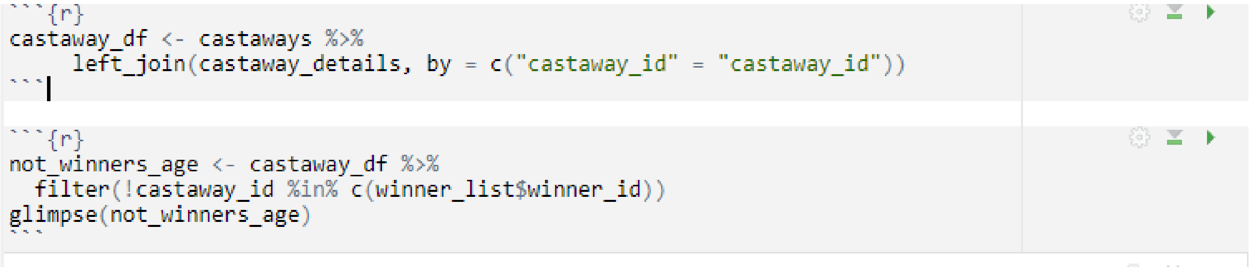
The following is how the null and alternative hypothesis is formulated:

Null_Hypothesis: Age does not have an impact on chances of winning.

Alternative_Hypothesis: Age does have an impact on chances of winning.

Calculating non-winners of survivor and their ages:

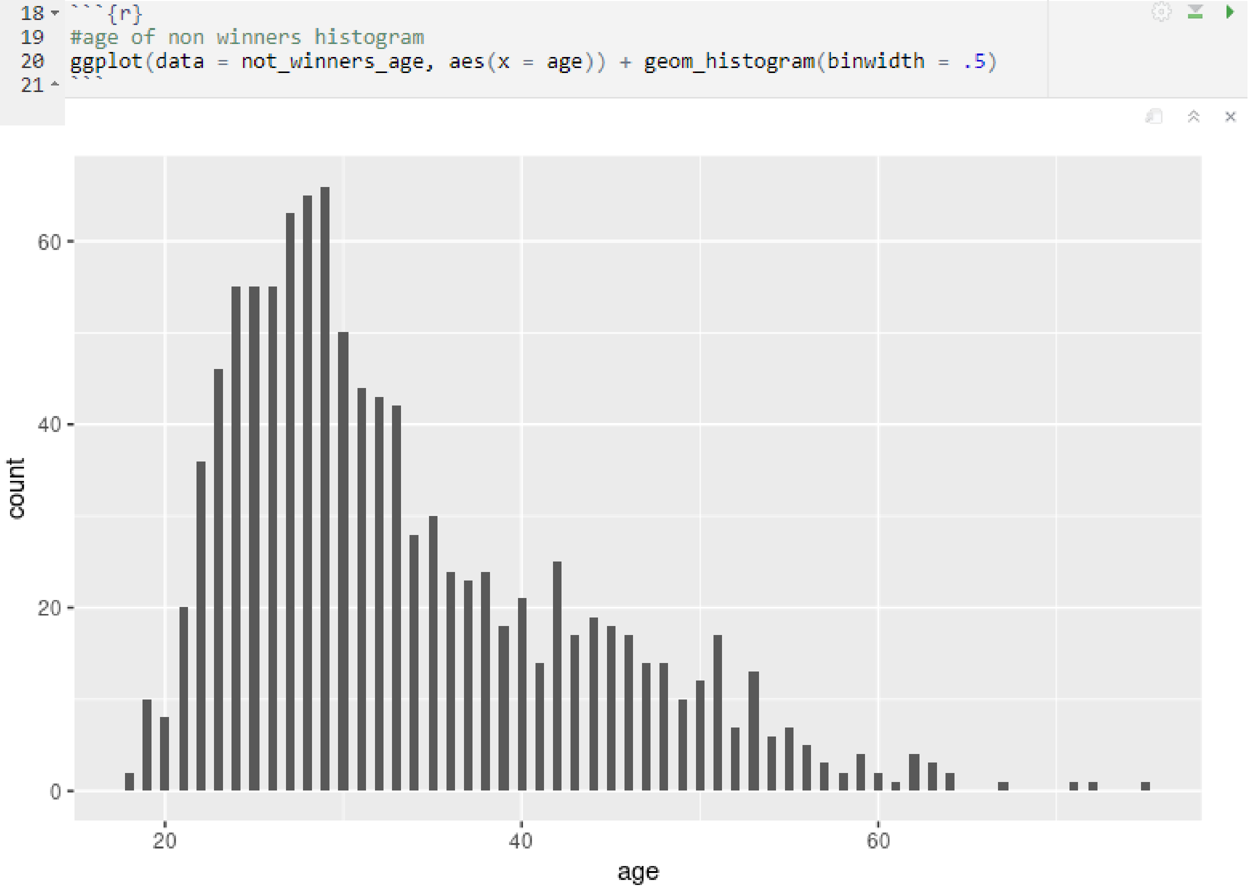
First, the data frame that isolated all contestants who did not win their season was created.



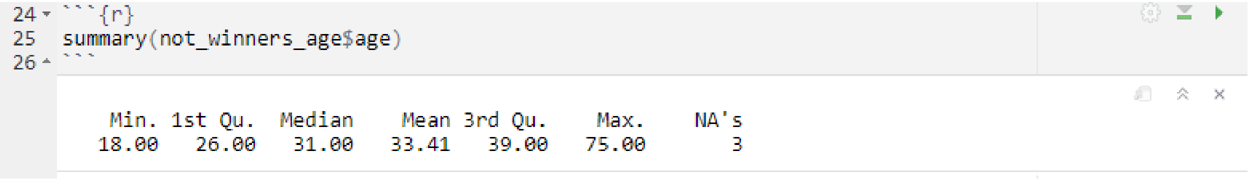
~~~
```{r}
castaway_df <- castaways %>%
     left_join(castaway_details, by = c("castaway_id" = "castaway_id"))
```
~~~

~~~
```{r}
not_winners_age <- castaway_df %>%
  filter(!castaway_id %in% c(winner_list$winner_id))
glimpse(not_winners_age)
```
~~~

Then, a histogram and a summary to show the distribution of age in non-winners was created.



~~~
```{r}
#age of non winners histogram
ggplot(data = not_winners_age, aes(x = age)) + geom_histogram(binwidth = .5)
```
~~~

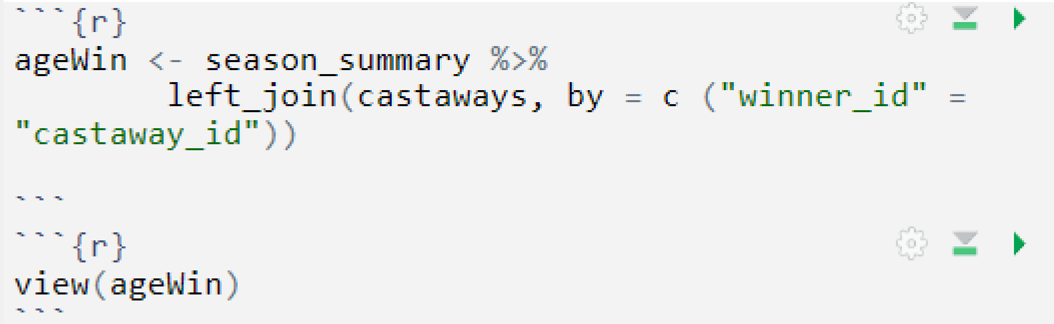


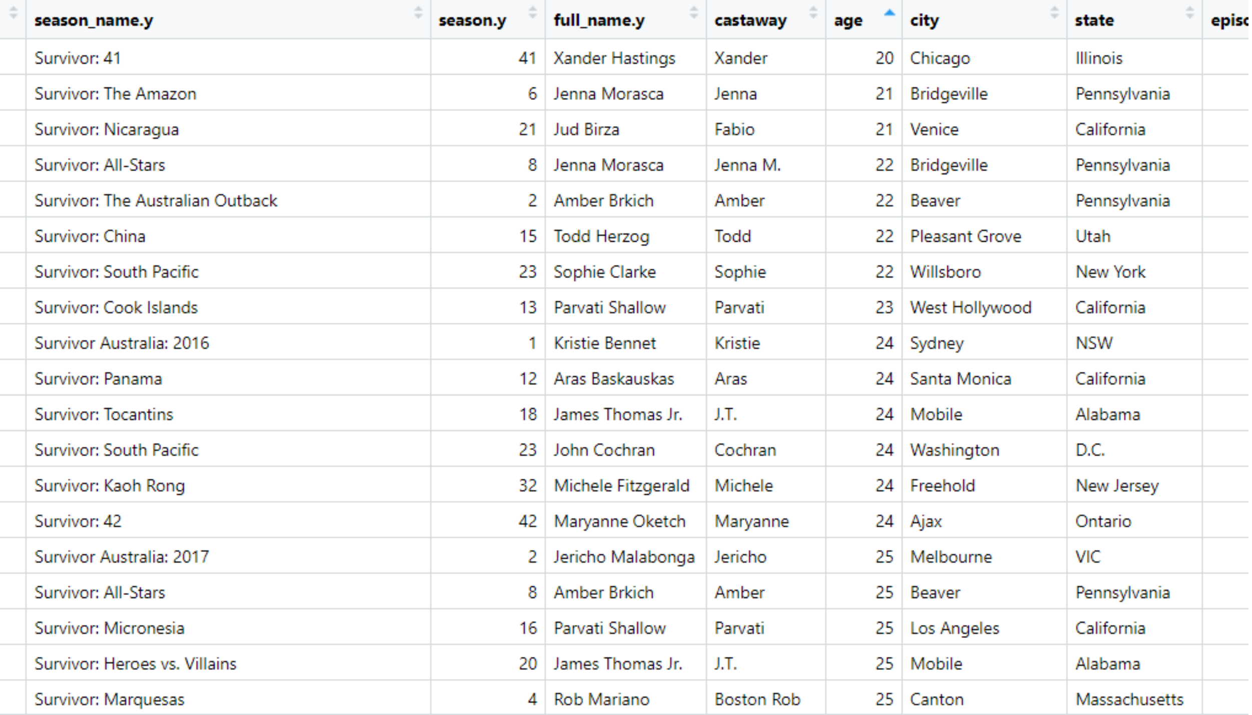
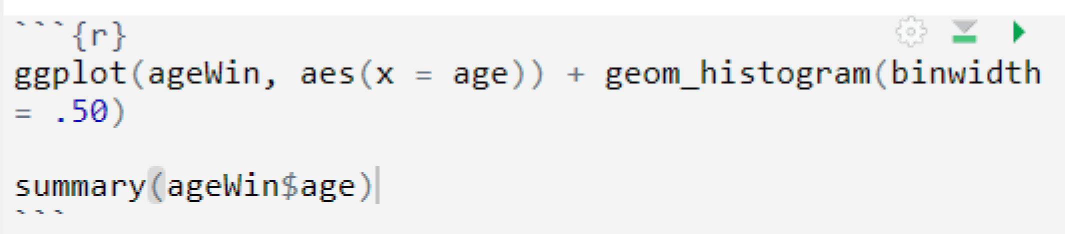
~~~
```{r}
summary(not_winners_age$age)
```
~~~

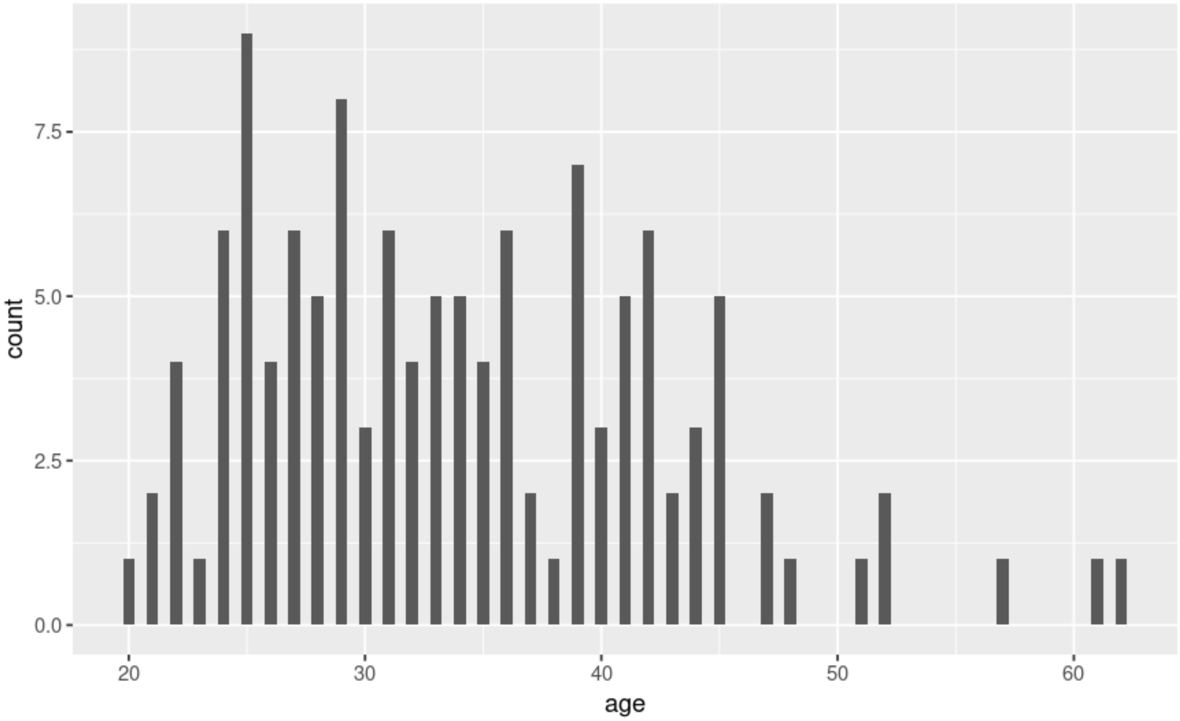
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  18.00   26.00   31.00   33.41   39.00   75.00       3 
```

After looking at this data, we can see the ages of non winning castaways is right skewed with the majority of castaways between 26 and 39 years old.

Calculating winners of survivor and their ages:
The first step was to look at the data frame and plot all the data in columns to look at all the information. Then, the different variables were called in order to graph multiple variables together correctly.



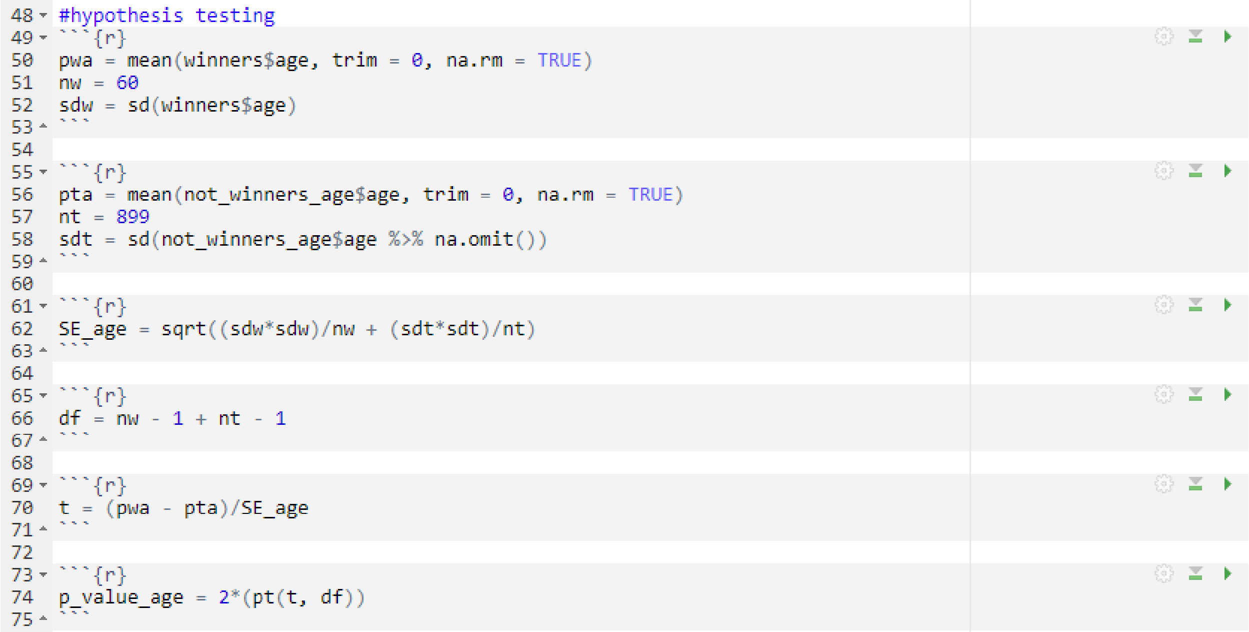
~~~
```{r}
ageWin <- season_summary %>%
        left_join(castaways, by = c ("winner_id" =
"castaway_id"))

~~~
```{r}
view(ageWin)
```
```



	season_name.y	season.y	full_name.y	castaway	age	city	state	episo
	Survivor: 41	41	Xander Hastings	Xander	20	Chicago	Illinois	
	Survivor: The Amazon	6	Jenna Morasca	Jenna	21	Bridgeville	Pennsylvania	
	Survivor: Nicaragua	21	Jud Birza	Fabio	21	Venice	California	
	Survivor: All-Stars	8	Jenna Morasca	Jenna M.	22	Bridgeville	Pennsylvania	
	Survivor: The Australian Outback	2	Amber Brkich	Amber	22	Beaver	Pennsylvania	
	Survivor: China	15	Todd Herzog	Todd	22	Pleasant Grove	Utah	
	Survivor: South Pacific	23	Sophie Clarke	Sophie	22	Willsboro	New York	
	Survivor: Cook Islands	13	Parvati Shallow	Parvati	23	West Hollywood	California	
	Survivor Australia: 2016	1	Kristie Bennet	Kristie	24	Sydney	NSW	
	Survivor: Panama	12	Aras Baskauskas	Aras	24	Santa Monica	California	
	Survivor: Tocantins	18	James Thomas Jr.	J.T.	24	Mobile	Alabama	
	Survivor: South Pacific	23	John Cochran	Cochran	24	Washington	D.C.	
	Survivor: Kaoh Rong	32	Michele Fitzgerald	Michele	24	Freehold	New Jersey	
	Survivor: 42	42	Maryanne Oketch	Maryanne	24	Ajax	Ontario	
	Survivor Australia: 2017	2	Jericho Malabonga	Jericho	25	Melbourne	VIC	
	Survivor: All-Stars	8	Amber Brkich	Amber	25	Beaver	Pennsylvania	
	Survivor: Micronesia	16	Parvati Shallow	Parvati	25	Los Angeles	California	
	Survivor: Heroes vs. Villains	20	James Thomas Jr.	J.T.	25	Mobile	Alabama	
	Survivor: Marquesas	4	Rob Mariano	Boston Rob	25	Canton	Massachusetts	

Then after looking at the data set to see the ages and the castaway player's variables names. By using a histogram to plot the data, it was easier to look specifically at the age and who had won that season. Then the average age for the winner of Survivor was calculated in order to see what the most common winner's age was and to look at the median to see if there was any slight data skewing. There is a slight skew, but overall the average Survivor winner is in their 30s.



```
```{r}
ggplot(ageWin, aes(x = age)) + geom_histogram(binwidth
= .50)

summary(ageWin$age)
```
```
```



Min	1st Qu.	Median	Mean	3rd Qu.	Max.
20.00	27.00	33.00	34.07	40.00	62.00

Based on these graphs and tables we can see that overall the average winners tend to be on the younger side, even though we have a couple of outliers that are older than 45.

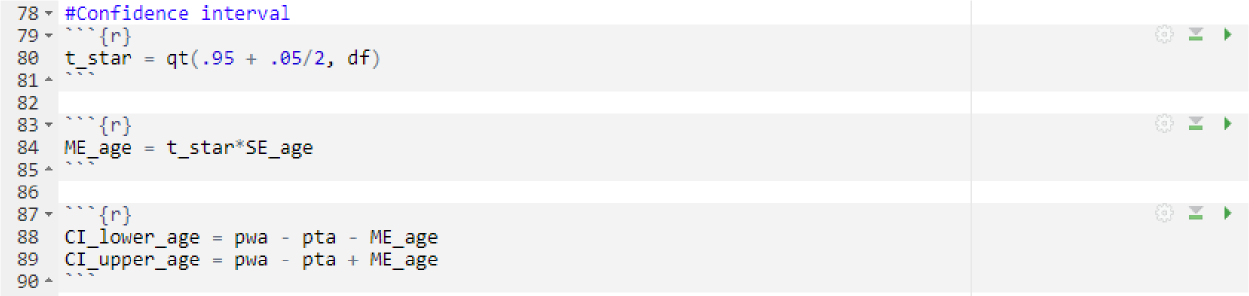
To test the hypothesis (Does the rate of challenge wins affect your chances of winning?) The following is how the null and alternative hypothesis was formatted:
Null_Hypothesis: Age does not have an impact on chances of winning.
Alternative_Hypothesis: Age does have an impact on chances of winning.
To test this hypothesis, an inference for the difference of 2 means test was used as shown below.



~~~r
#hypothesis testing
```{r}
pwa = mean(winners$age, trim = 0, na.rm = TRUE)
nw = 60
sdw = sd(winners$age)
```

```{r}
pta = mean(not_winners_age$age, trim = 0, na.rm = TRUE)
nt = 899
sdt = sd(not_winners_age$age %>% na.omit())
```

```{r}
SE_age = sqrt((sdw*sdw)/nw + (sdt*sdt)/nt)
```

```{r}
df = nw - 1 + nt - 1
```

```{r}
t = (pwa - pta)/SE_age
```

```{r}
p_value_age = 2*(pt(t, df))
```
~~~

This resulted in a p-value of .4101. A confidence level of 95% was chosen, which means alpha is .05. Because the p-value is higher than alpha, we fail to reject the null hypothesis. Overall, this means that the distribution of ages amongst the winners is approximately equal to the distribution of the ages among the non-winners.

To check for errors, a confidence interval was constructed, as seen below.



~~~r
#Confidence interval
```{r}
t_star = qt(.95 + .05/2, df)
```

```{r}
ME_age = t_star*SE_age
```

```{r}
CI_lower_age = pwa - pta - ME_age
CI_upper_age = pwa - pta + ME_age
```
~~~

The confidence interval is (-3.1881, 1.2998). This means that there is a 95% chance that the difference between the two means is between -3.1881 and 1.2998. This confirms our acceptance of the null hypothesis.

Does Personality Type affect your chances of winning?

To test the hypothesis:

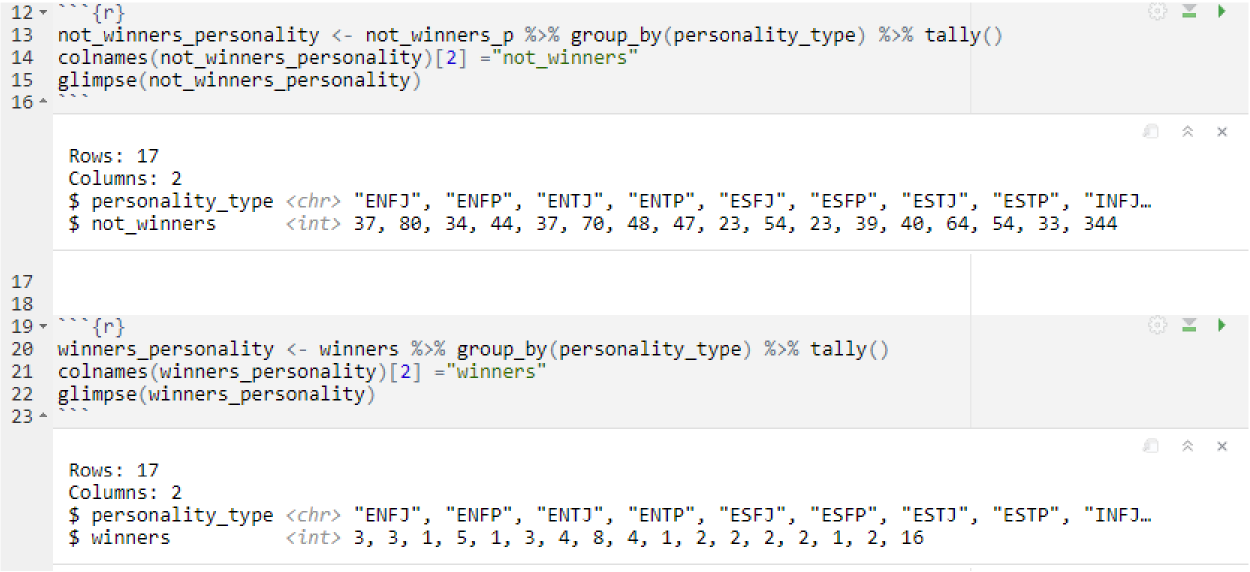
The following is how the null and alternative hypothesis are formatted:

Null_Hypothesis: Personality type does not impact chances of winning.

Alternative_Hypothesis: Personality type does have an impact on chances of winning.

Creating the data frame:

First, was to find the number of each personality type, in both winners and non winners.



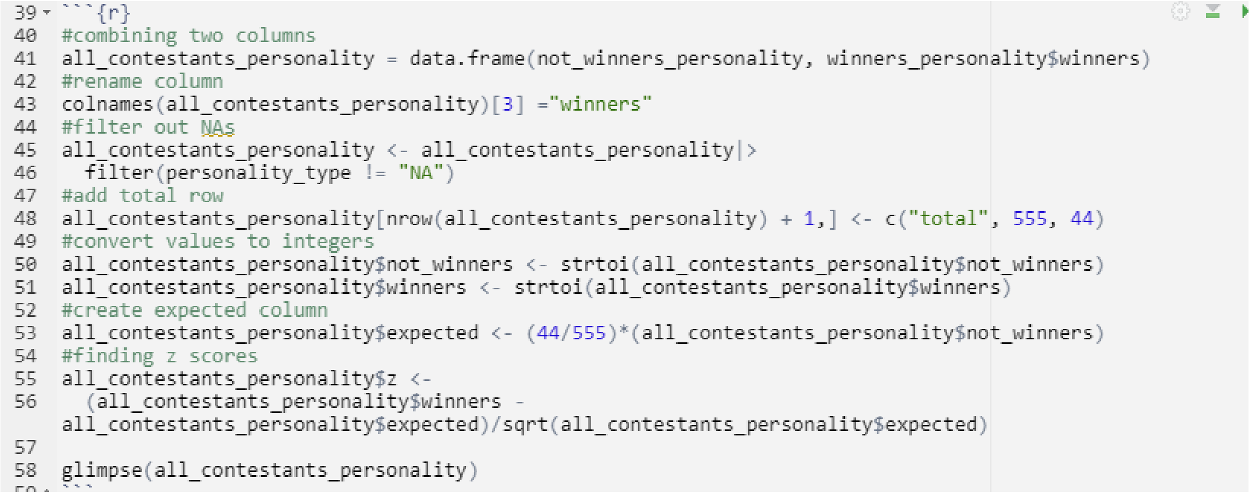
~~~
```{r}
not_winners_personality <- not_winners_p %>% group_by(personality_type) %>% tally()
colnames(not_winners_personality)[2] ="not_winners"
glimpse(not_winners_personality)
```
~~~

```
Rows: 17
Columns: 2
$ personality_type <chr> "ENFJ", "ENFP", "ENTJ", "ENTP", "ESFJ", "ESFP", "ESTJ", "ESTP", "INFJ…
$ not_winners      <int> 37, 80, 34, 44, 37, 70, 48, 47, 23, 54, 23, 39, 40, 64, 54, 33, 344
```

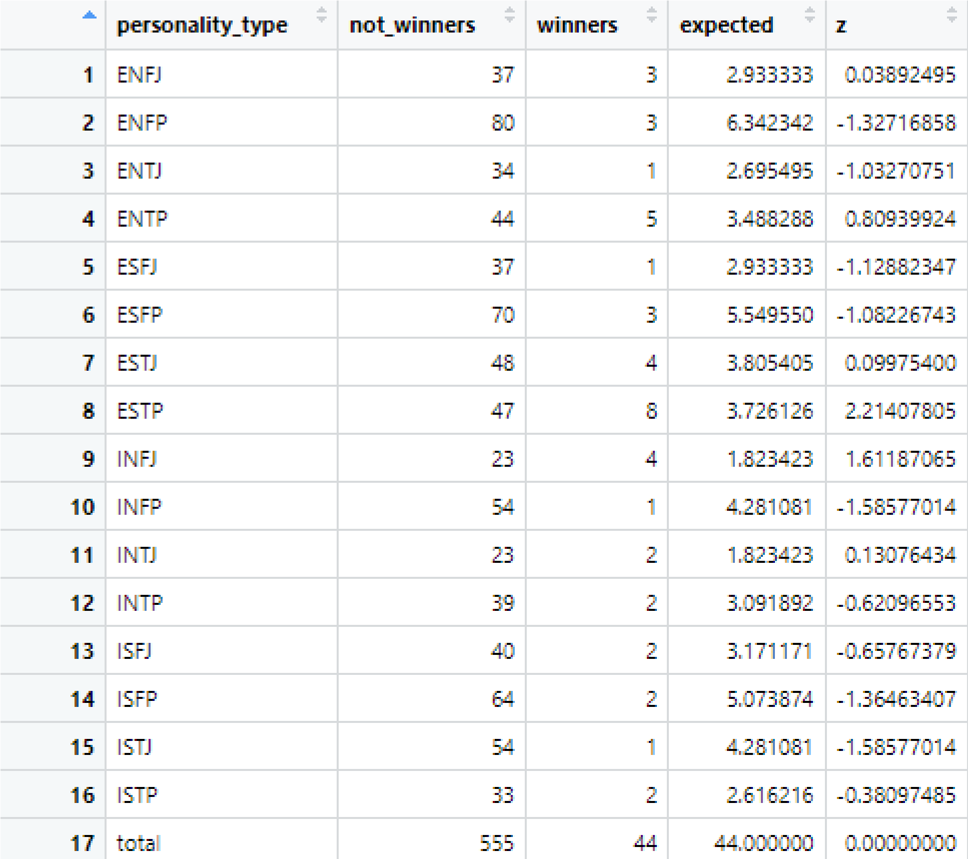
~~~
```{r}
winners_personality <- winners %>% group_by(personality_type) %>% tally()
colnames(winners_personality)[2] ="winners"
glimpse(winners_personality)
```
~~~

```
Rows: 17
Columns: 2
$ personality_type <chr> "ENFJ", "ENFP", "ENTJ", "ENTP", "ESFJ", "ESFP", "ESTJ", "ESTP", "INFJ…
$ winners          <int> 3, 3, 1, 5, 1, 3, 4, 8, 4, 1, 2, 2, 2, 2, 1, 2, 16
```

Next, the two tables were combined. An additional column was created to compute the expected number of winners using the equation non-winner personality value * total number of winners / total number of participants. Finally, to compare the expected versus actual number of winners per personality type, a z score was computed using the equation (winner personality value - expected winner personality value) / square root (winner personality value). All of these steps and the resulting table is shown below.

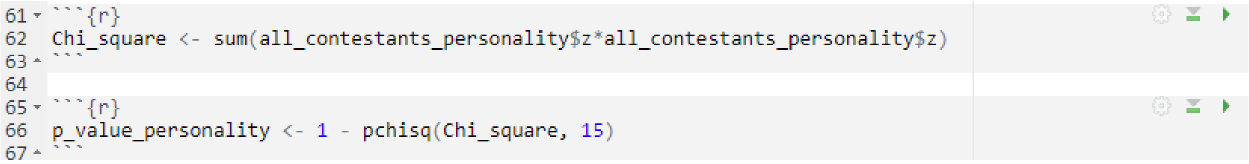


~~~
```{r}
#combining two columns
all_contestants_personality = data.frame(not_winners_personality, winners_personality$winners)
#rename column
colnames(all_contestants_personality)[3] ="winners"
#filter out NAs
all_contestants_personality <- all_contestants_personality|>
  filter(personality_type != "NA")
#add total row
all_contestants_personality[nrow(all_contestants_personality) + 1,] <- c("total", 555, 44)
#convert values to integers
all_contestants_personality$not_winners <- strtoi(all_contestants_personality$not_winners)
all_contestants_personality$winners <- strtoi(all_contestants_personality$winners)
#create expected column
all_contestants_personality$expected <- (44/555)*(all_contestants_personality$not_winners)
#finding z scores
all_contestants_personality$z <-
  (all_contestants_personality$winners -
all_contestants_personality$expected)/sqrt(all_contestants_personality$expected)

glimpse(all_contestants_personality)
```
~~~



	personality_type	not_winners	winners	expected	z
1	ENFJ	37	3	2.933333	0.03892495
2	ENFP	80	3	6.342342	-1.32716858
3	ENTJ	34	1	2.695495	-1.03270751
4	ENTP	44	5	3.488288	0.80939924
5	ESFJ	37	1	2.933333	-1.12882347
6	ESFP	70	3	5.549550	-1.08226743
7	ESTJ	48	4	3.805405	0.09975400
8	ESTP	47	8	3.726126	2.21407805
9	INFJ	23	4	1.823423	1.61187065
10	INFP	54	1	4.281081	-1.58577014
11	INTJ	23	2	1.823423	0.13076434
12	INTP	39	2	3.091892	-0.62096553
13	ISFJ	40	2	3.171171	-0.65767379
14	ISFP	64	2	5.073874	-1.36463407
15	ISTJ	54	1	4.281081	-1.58577014
16	ISTP	33	2	2.616216	-0.38097485
17	total	555	44	44.000000	0.00000000

To test the hypothesis (Does your personality type affect your chances of winning?) The following is how the null and alternative hypothesis was formatted:
Null_Hypothesis: Personality type does not have an impact on chances of winning.
Alternative_Hypothesis: Personality type does have an impact on chances of winning.
To test this hypothesis, a p_value was constructed:



~~~
```{r}
Chi_square <- sum(all_contestants_personality$z*all_contestants_personality$z)
~~~

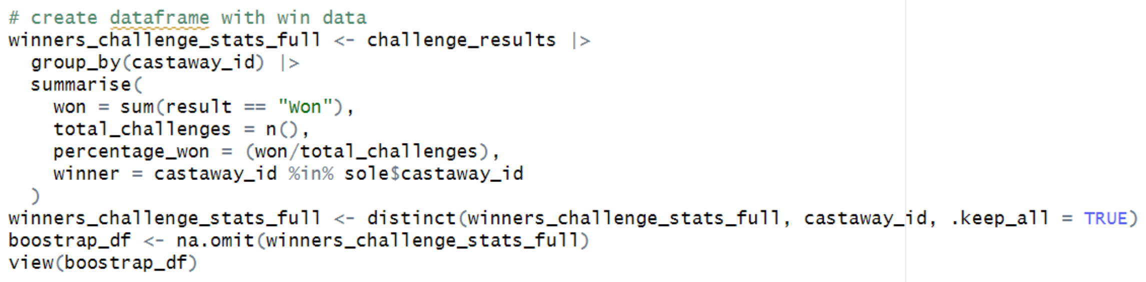
```{r}
p_value_personality <- 1 - pchisq(Chi_square, 15)
```
```

This resulted in a p-value of .1271 which is high enough to suggest the null hypothesis is true.
```

Results & Discussion

Does the rate of challenge wins affect your chances of winning?

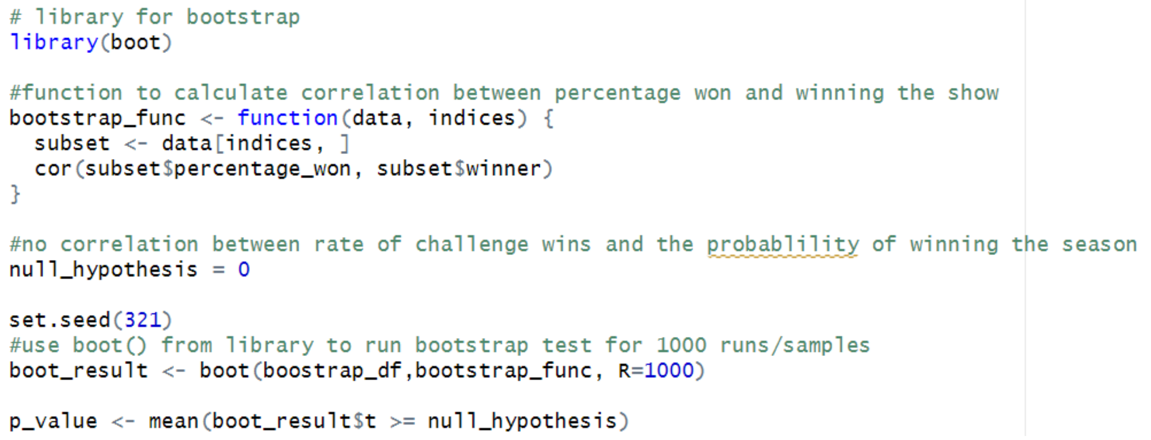
The average rate of challenge wins for contestants is not significantly different from the rate of challenge wins for season winners. Therefore, the rate of challenge wins does not affect your chances of winning.

In order to validate the findings of the hypothesis, the utilization of bootstrapping was used on a dataframe of challenge wins and season winners. The following code showcases how the data frame was produced:



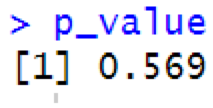
```
# create dataframe with win data
winners_challenge_stats_full <- challenge_results |>
  group_by(castaway_id) |>
  summarise(
    won = sum(result == "Won"),
    total_challenges = n(),
    percentage_won = (won/total_challenges),
    winner = castaway_id %in% sole$castaway_id
  )
winners_challenge_stats_full <- distinct(winners_challenge_stats_full, castaway_id, .keep_all = TRUE)
boostrap_df <- na.omit(winners_challenge_stats_full)
view(boostrap_df)
```

The library of "boot" is utilized to run the bootstrap validation on the created dataframe. In order to utilize the boot() function, a bootstrap_func() was created to get the correlation between win percentage and the probability of winning the show. The following code is how the boot() function was run with using the data frame created and the boostrap_func() utilized:



```
# library for bootstrap
library(boot)

#function to calculate correlation between percentage won and winning the show
bootstrap_func <- function(data, indices) {
  subset <- data[indices, ]
  cor(subset$percentage_won, subset$winner)
}

#no correlation between rate of challenge wins and the probablility of winning the season
null_hypothesis = 0

set.seed(321)
#use boot() from library to run bootstrap test for 1000 runs/samples
boot_result <- boot(boostrap_df,bootstrap_func, R=1000)

p_value <- mean(boot_result$t >= null_hypothesis)
```

Once the code was run, a p_value was found:



```
> p_value
[1] 0.569
```

With an alpha of 0.05 used, the p_value found is greater than the alpha (0.569 > 0.05). Thus, we do not have sufficient evidence to reject the null hypothesis which states there is no correlation between number of challenges won and the probability of winning the season.

If you are a returning player, are you more likely to win the game?

Since the p-value found was less than the alpha, we reject the null hypothesis and accept the alternative hypothesis, as there is significant evidence to support the claim that being a returning player does have an effect on your chances of winning. In this case, returning players are more likely to win the competition than those that have never been on the show before.

This means, in order to create the best participant with the highest chance of winning the show, one characteristic that should be included to increase chances is for the participant to be a returning player, as they have a 13.33% higher chance of winning the game.

Does your age impact your chances of winning?

Based on our findings, we fail to reject our null hypothesis. So age does not have an impact on the chances of winning, as shown in our hypothesis testing. If age mattered, we would have a noticeable difference in the distribution of ages among the winners as compared to those who did not win. So based on this data, age does not have an impact on the chances of winning.

Does Personality Type affect your chances of winning?

While hypothesis testing may suggest that personality type has little impact on chances of winning, we can also take a closer look at individual z scores to get a better understanding of how having specific personality types could impact your chances of winning.
For example, the ESTP personality type has a high z score. This means that more people from this personality type won than expected. People with this personality type are characterized as energetic, observant, impulsive, and competitive. This personality type is also known as the "entrepreneurs" with some famous examples, including Eddie Murphy, Madonna, and Ant-Man. Another personality type with a high z-score is INFJ. People with this personality type are characterized as introverted, driven, compassionate, and have a strong sense of morals. This personality type is also known as the "advocate," with some famous examples including Marie Kondo, Lady Gaga, and Rose from Titanic.

In contrast, INFP had a low z-score suggesting an unexpectedly small number of people from this personality type won. People with this personality type are characterized as idealistic, empathetic, curious, and have a strong sense of self. This personality type is also known as the "mediator," with some famous examples including Björk, Tom Hiddleston, and Frodo Baggins. Another personality type with a low z-score is ISTJ. People with this personality type are characterized as dedicated, structured, honest, and practical. This personality type is also known as the "logistician," with some famous examples including Natalie Portman, George Bush, and Hermione Granger.

Conclusions

Overall, we decided to converge and investigate these three different hypotheses because they are the most talked about aspects of winning the game Survivor. We believe that these hypotheses will lead us to a better understanding of how to win Survivor from a statistical aspect, along with finding other outcomes along the way. From this, we wanted to determine if this was actually true and discover potential explanations for why these three parts of the game are so important. We also took into account the social aspect of the game by choosing to analyze if being a returning player impacted one's chances of winning the game. We were wondering if coming back with alliances/relationships formed outside of the game, plus having the advantage of already playing the game, helps one win the game again or hurt their chances by being too strong of a player and, in turn, a threat.

We have seen from our initial testing that two of the three variables from our hypothesis, being a returning player and your age, does have an impact on your chances of winning Survivor. The number of challenge wins, however, does not impact your chances of winning the overall season. After our testing, we did indeed see that being a returning player does have an impact on winning Survivor and allows for an advantage by already having knowledge of how to play the game. Along with knowing a good social game plus forming alliances well in the game and having built stronger ones outside, give these castaways a leg up in Survivor, by having a 13.33% higher chance of winning.

We then looked at age as a winning factor, and if a person is around the age of 33, they are a target age to win Survivor. By this, we can see that castaways playing this game are at a disadvantage by being younger than 27 and older than 35. We came to an initial conclusion and believed that age did have an impact on winning the game. However, we did some more robust testing, and we actually discovered that compared to the age distribution for those who won was approximately equal to the distribution of those who did not win. This means that age is not a factor in winning. Because of this, we decided to test to see if personality type is a significant factor in winning.

Based on our findings for personality type, our data shows that while one's personality type is not a significant factor in winning, people with a more extroverted personality tend to win more. It is interesting that our data showed that personality did not have a large impact because, based on outside research from different articles, the most important winning factor is indeed one's personality or overall social game. A castaway's personality impacts how they play the game. If their personality is too dominating, then people will use it as leverage to vote them out, compared to others who are friendly and show who they are as a person. While there are multiple personality types as well, if you tend to be a more introverted person, this personality would hurt you by not taking enough big moves and then allowing you to stay in the game because you will not get voted for in the end when it matters. The social aspect of the game seems to be the more important part of winning because that is how one controls the vote and then gets the jury to vote for them at the end of the game (Lindbergh).

There are indeed some limitations that we have discovered so far, as well. While our sample size is a good size, the overall pool for Survivor winners is only 40 people. The bigger the sample pool, the more accurate findings could be discovered to allow for better conclusions. We have also seen from our data analysis that only using a few statistical techniques are beneficial for discovering general findings and then conclusions, but to make our analysis more definitive.

Overall what we have concluded from our data analysis is that from our hypotheses is that while your age and potentially personality might not matter, being a returning player and the rate of challenge wins does indeed impact chances of being a Survivor winner.

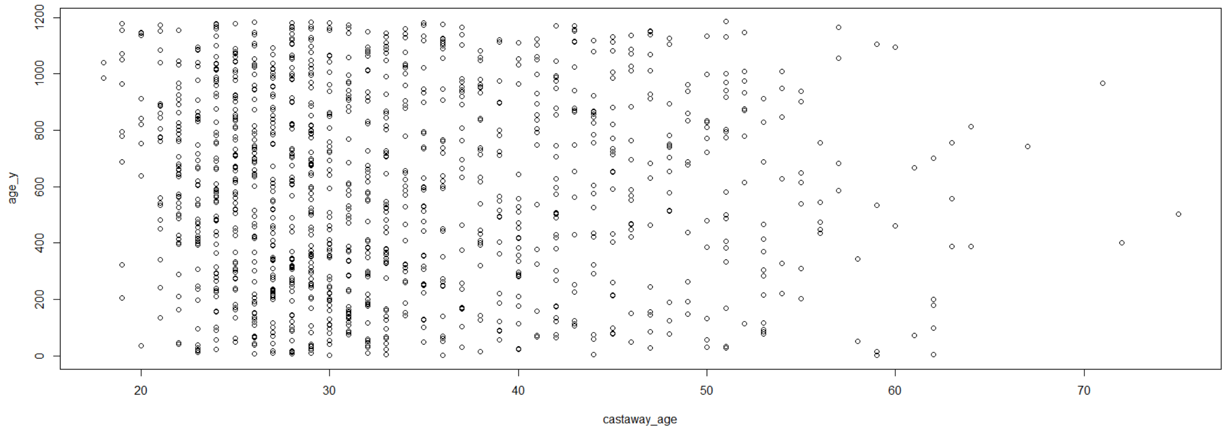
Bibliography

Lindbergh, Ben. “How Hardcore Edit Analysts Unravel the Secrets of 'Survivor'.” *The Ringer*, The Ringer, 13 May 2020, https://www.theringer.com/tv/2020/5/13/21256821/survivor-edgic-analysis-winners-edit.

Appendices

Appendix A

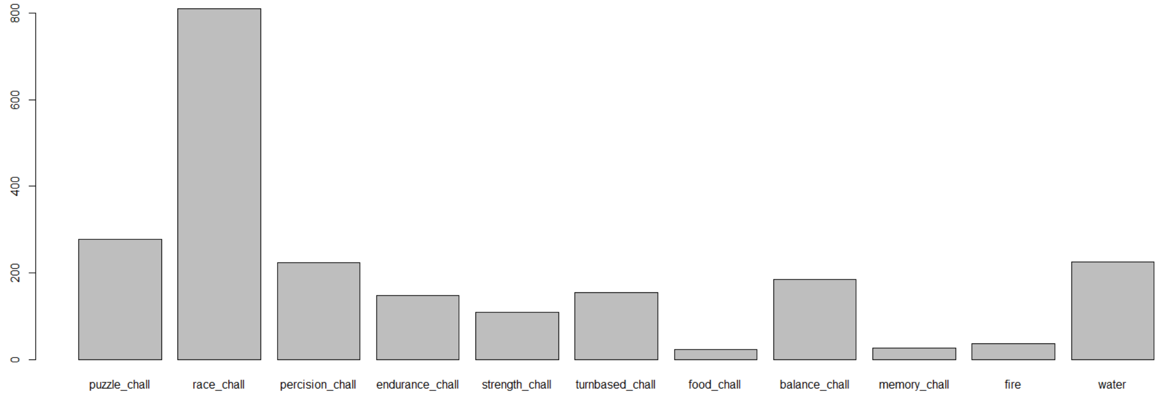
Appendix A1



Appendix A2

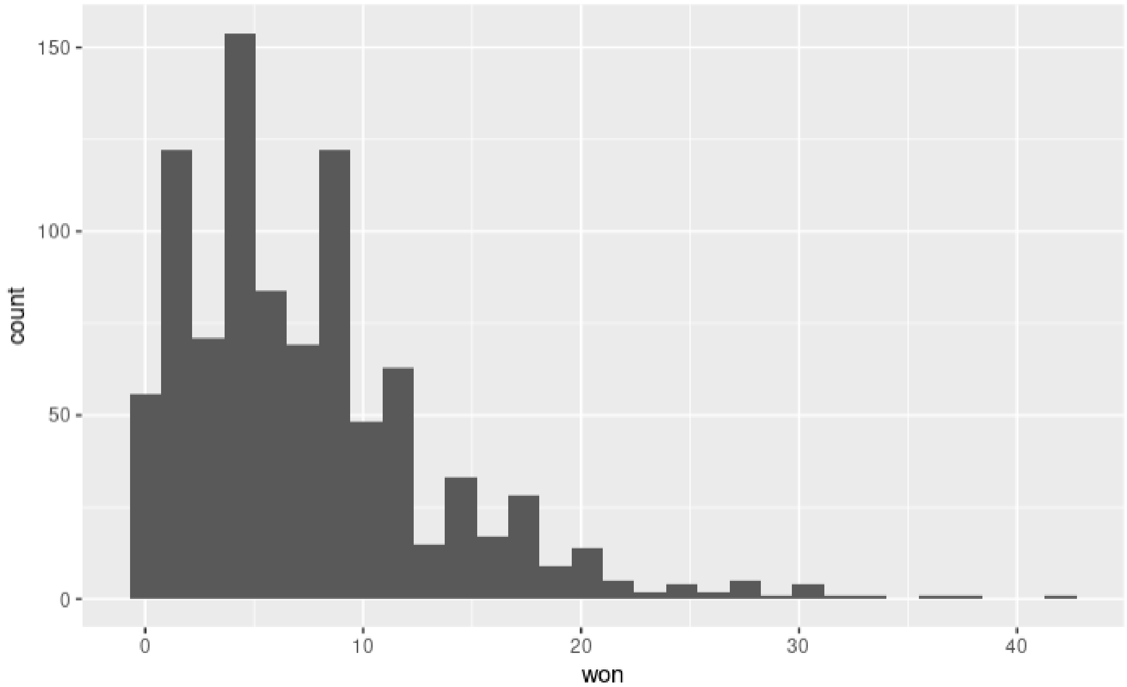
Min	Q1	Median	Mean	Q3	Max	NA's		
18	26	31	33.43	39	75	3		

Appendix B



Appendix C

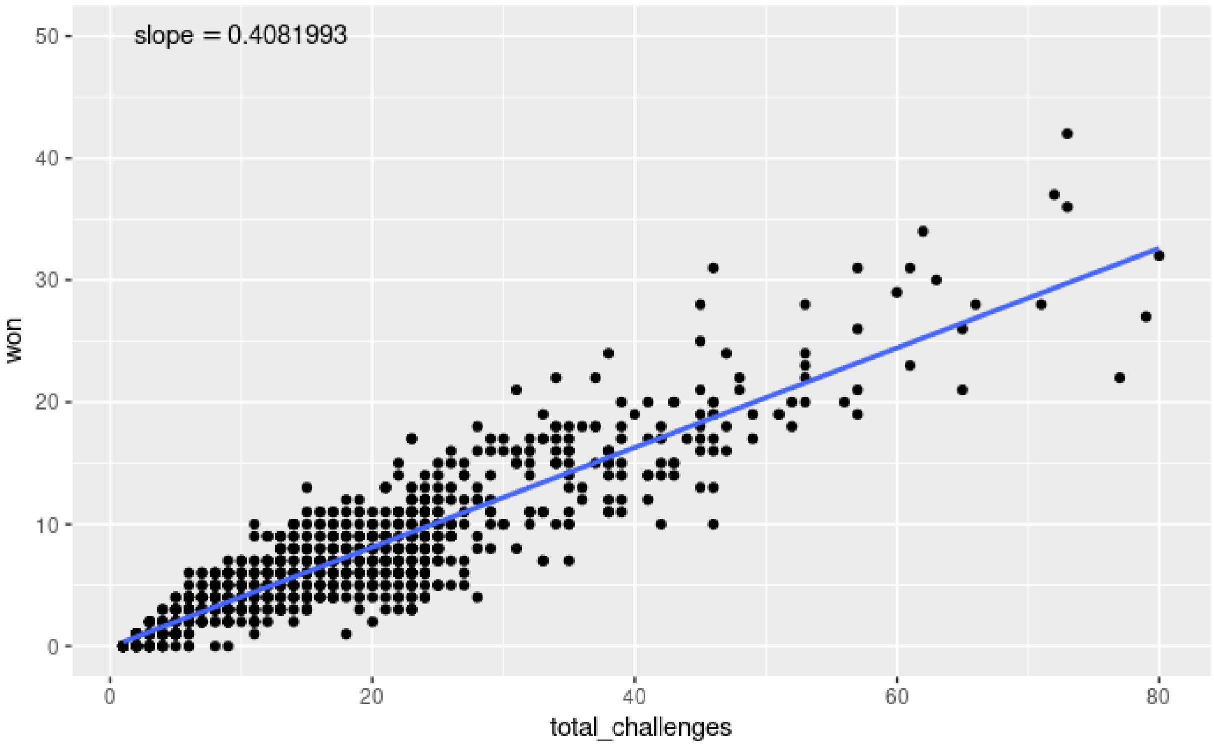
Appendix C1



Appendix C2

Min	Q1	Median	Mean	Q3	Max	NA
0	3	6	7.46	10	42	24

Appendix D


slope = 0.4081993
won
total_challenges
0
10
20
30
40
50
0
20
40
60
80