SDS 291 - Fat Bear 2020-2021

Laura Mora, Mariem Snoussi, Maya Crandall-Malcolm

2021-12-07

Contents

```
library(tidyverse)
library(openintro)
library(mosaic)
library(modeest)
library(GGally)
```

```
bears<-read.csv("FullBearData.csv")
survey20<-read.csv("Survey2020.csv")
survey21<-read.csv("Survey2021.csv")
```

Purpose

The purpose of this analysis is to evaluate what factors play a major role in the winner of Fat Bear Week. Fat Bear Week is an online single elimination tournament held in Katmai National Park, Alaska. This year it was held from September 29th to October 5th. According to the website, where the voting takes place, who to vote for as the winner is subjective. Some voters may choose to vote for a particular bear based on dramatic change in weight over the year. Other voters might choose to vote based on additional challenges the bears may face in order to gain weight. But in general the idea is to vote for the fattest bear, in the voter's opinion. We hypothesize that the more angled away from the camera the bear is, i.e. the closer the butt of the bear is to the camera, the more votes that bear will receive. We also hypothesize that the probability that a bear will win the second round increases as the number of votes that bear obtains in the first round increases. That is, the likelihood of success, where success is that a particular bear wins in the second round, increases as the number of votes the bear won in the first round increases. We will also explore other variables that might play a role in the who named winner of Fat Bear Week; such as, fur color, change in size, and more.

Data

We plan to use data about the last two years of the Fat Bear Week competition that we compiled ourselves from various sources (Fat Bear Week Voting Page, Katmai Park publications, Katmai Bearcams Fan Wiki). We will at some point link to our dataset. There are many more bears at Katmai National Park than are featured in the Fat Bear Week competition. Former park ranger Mike Fitz elaborated on the selection of each year's competitors, saying, < the bears are typically chosen by the park rangers, based on a variety of factors such as their time of arrival at Brooks River. An early summer arrival at the river ensures that the bear is photographed before and after their weight gain. A compelling storyline is also a driving factor when choosing a contestant. The audience loves to see a veteran bear with skill and resilience, like Otis, competing year after year > [1]. Although the contest began in 2014, we restricted our investigation to the contests that took place in 2020 and 2021. Contests prior to those years took place on Facebook with the bear whose post received the most "Likes" winning each round, where we have no way to control for bias due to the Facebook algorithm for amplifying posts given previous likes. In 2020 and 2021, voting took place on a designated webpage [2] which guarantees that all respondents saw both bears for each match-up they voted in.

Respondent perception of a bear's size, color, and position in relation to the camera is subjective and vary between respondents. To determine the attributes size (before & after), color, and angle of each bear relative to the camera, we conducted two brief surveys (one for each contest year) with human respondents. The survey asked respondents to look at the official before and after photos of the bears from the contests, and report each bear's size, angle in relation to the camera, and fur color. We used convenience sampling of primarily our friends and family. We obtained the results of the survey for each bear by identifying the most frequent response for the questions related to fur color and size, and took the mean of the responses for angle to the camera.

Population

Each row of the dataset is one half of a matchup in the tournament. Each row contains the bear's designated number, name (if applicable), contest year, round, opponent's designated number, number of votes received in that matchup, outcome of the matchup, outcome of the whole tournament, number of previous Fat Bear Week wins, sex, age, fur color, summer size (before), fall size(after), and angle to the camera. It is our hope that our findings could be applied to predicting the winner of future Fat Bear Week tournaments.

Outcome Variable

The outcome variable for the first hypothesis is Winning Round 2. It's a binary response with 0 represnting a loss and 1 representing a win. The outcome variable for the second hypothesis is number of votes obtained in the first round. This is a quantitative variable in the positive integers, each vote will count as one unit.

Explanatory Variable

The explanatory variables are: Sex, a binary variable with values "M" for male and "F" for female. Color is a categorical variable. It's possible values are "Blond", "Light Brown", "Brown", "dark Brown" Age is a quantitative variable. Its unit is years. Presence of names is a binary variable. It will be 0 if the bear does not have a name and 1 when they have a name. Number of Votes is a quantitative variable, Each one point increase is equivalent to 1 additional vote towards the bear.

Exploratory Visualizations

A preliminary exploration of our data shows that there is some relationship between winning a round and some of our predictor variables using simple logistic regression. There appears to be positive associations

between angle to camera and the likelihood of winning a round, bear age and likelihood of winning a round, and number of votes and winning a round.
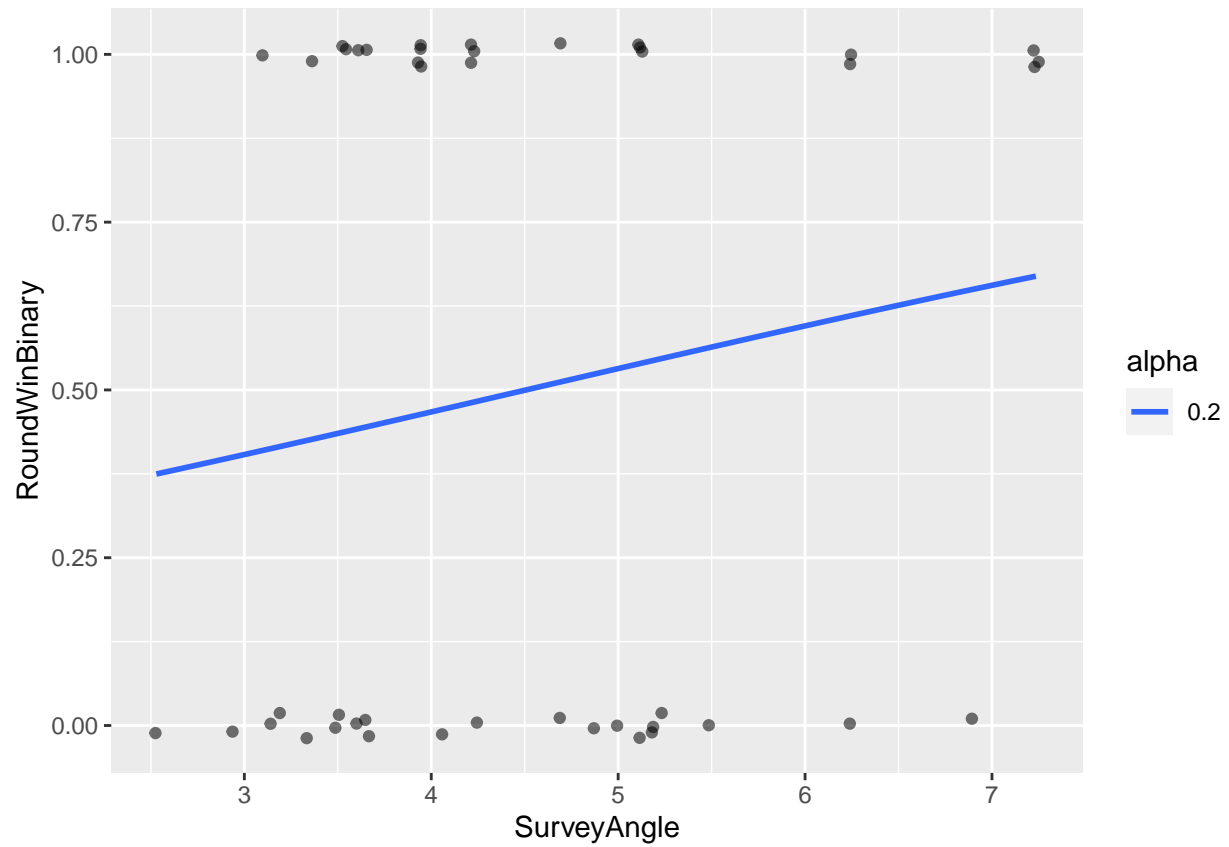
```
bears %>%
  select(RoundWinBinary, Sex, SurveyFur, SurveyBefore, SurveyAfter, SurveyAngle) %>%
  ggpairs()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
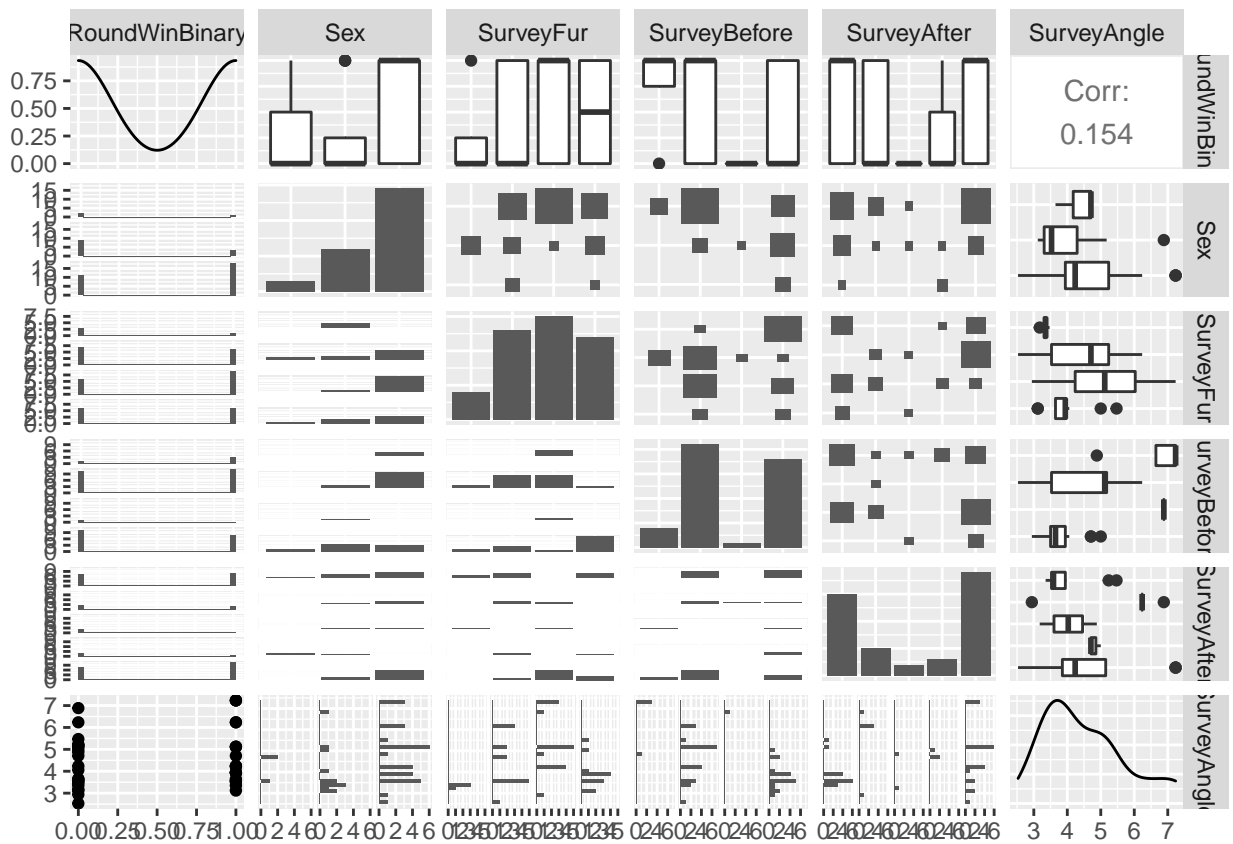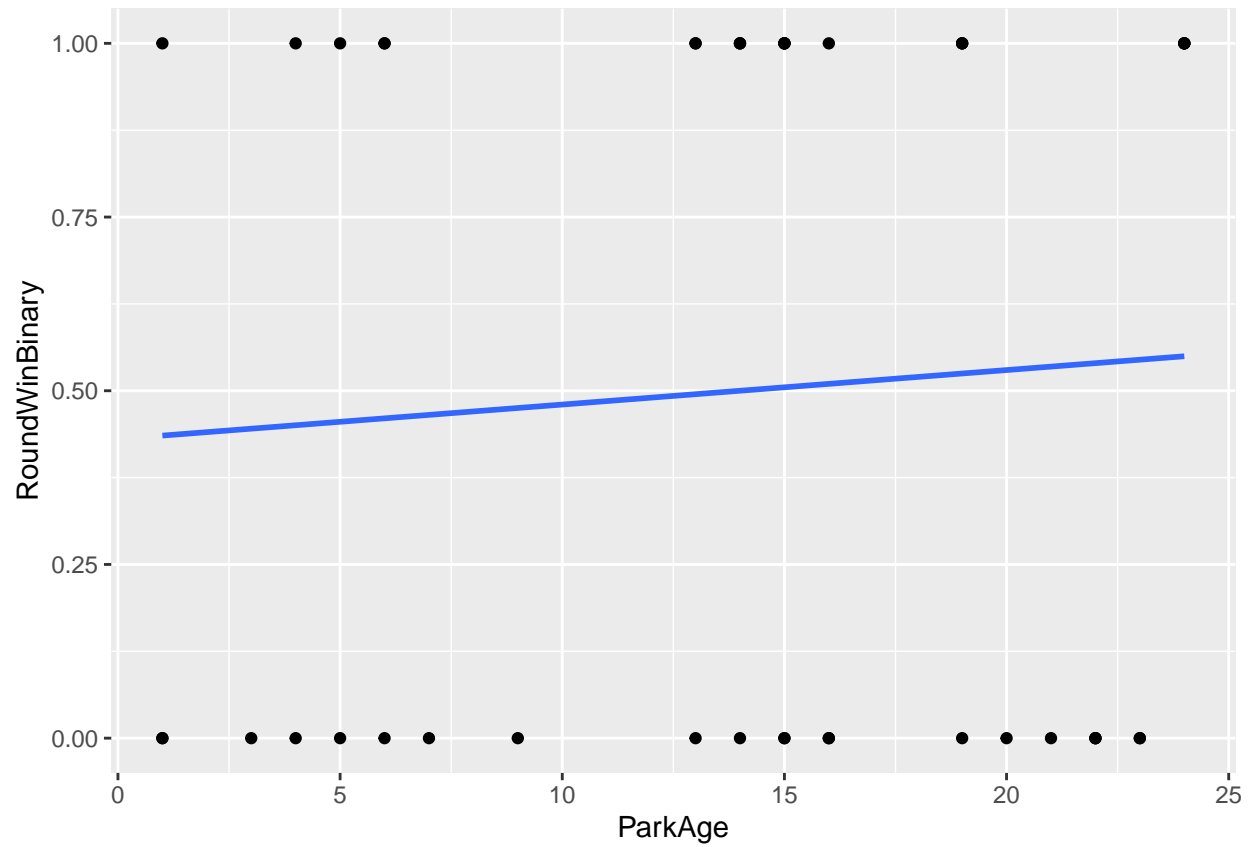


```
plot <- qplot(x = SurveyAngle, y = RoundWinBinary, data = bears,
              alpha = 0.2, show.legend = FALSE,
              geom = "jitter", height = 0.02, width = 0)
plot + geom_smooth(method = "glm",
method.args = list(family = "binomial"), se = 0)
```

```
## `geom_smooth()` using formula 'y ~ x'
```
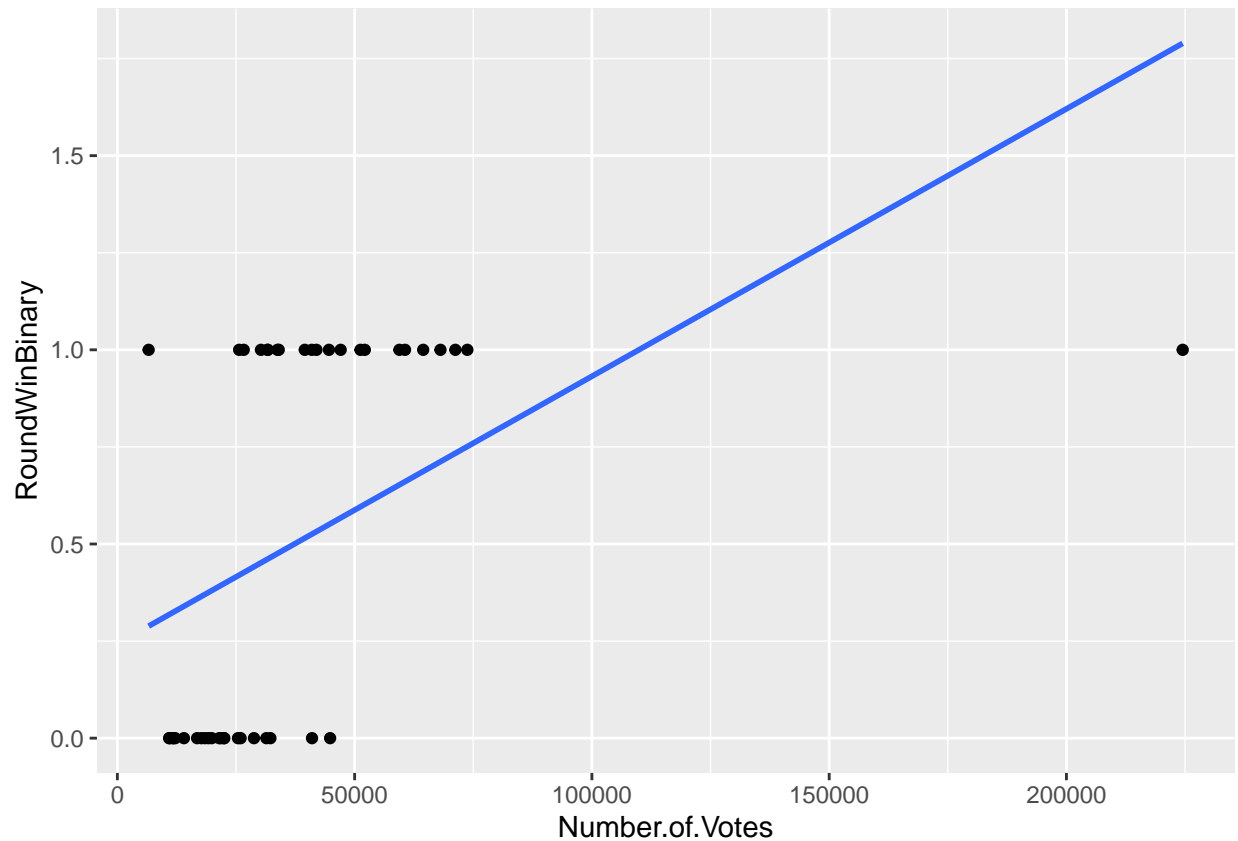
```
qplot(x=ParkAge, y=RoundWinBinary, data=bears) + geom_smooth(method="glm", se=FALSE)
```

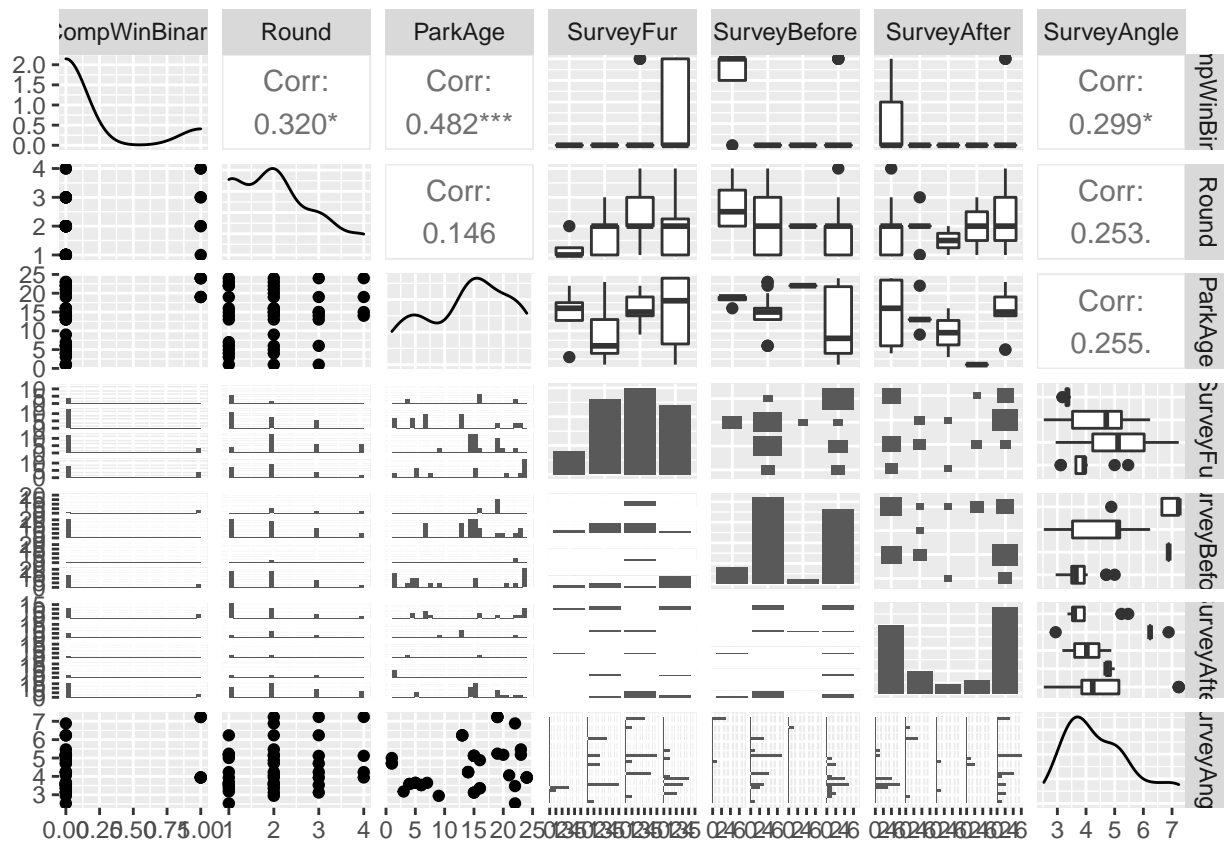## `geom_smooth()` using formula 'y ~ x'

```
qplot(x=Number.of.Votes, y=RoundWinBinary, data=bears) +
  geom_smooth(method="glm", se=FALSE)
```

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

```
bears %>%
  select(CompWinBinary, Round, ParkAge, SurveyFur, SurveyBefore, SurveyAfter, SurveyAngle) %>%
  ggpairs()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```r
# logit model for SurveyAngle and Win
logitAngle <- glm(CompWinBinary ~ SurveyAngle, data = bears, family = binomial)
summary(logitAngle)
```

```
##
## Call:
## glm(formula = CompWinBinary ~ SurveyAngle, family = binomial,
##     data = bears)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0165  -0.6395  -0.4181  -0.3618   2.1519
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.6499     1.7485  -2.659  0.00783 **
## SurveyAngle   0.6187     0.3316   1.866  0.06201 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.558  on 43  degrees of freedom
## Residual deviance: 34.907  on 42  degrees of freedom
## AIC: 38.907
##
```

```
## Number of Fisher Scoring iterations: 5
```

```r
#logit model with Win and Age
logitAge <- glm(CompWinBinary ~ ParkAge, data = bears, family = binomial)
summary(logitAge)
```

```
##
## Call:
## glm(formula = CompWinBinary ~ ParkAge, family = binomial, data = bears)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.26894  -0.30267  -0.18856  -0.02684   1.83616
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.5246     3.4931  -2.727   0.0064 **
## ParkAge       0.4234     0.1655   2.558   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.558  on 43  degrees of freedom
## Residual deviance: 23.357  on 42  degrees of freedom
## AIC: 27.357
##
## Number of Fisher Scoring iterations: 7
```

```r
logitAgeAngle <- glm(CompWinBinary ~ ParkAge*SurveyAngle, data = bears, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(logitAgeAngle)
```

```
##
## Call:
## glm(formula = CompWinBinary ~ ParkAge * SurveyAngle, family = binomial,
##     data = bears)
##
## Deviance Residuals:
##        Min         1Q     Median         3Q        Max
## -7.710e-05  -2.100e-08  -2.100e-08  -2.100e-08   3.658e-05
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1944.91  665486.13  -0.003    0.998
## ParkAge              85.61   29864.24   0.003    0.998
## SurveyAngle         309.93  108162.65   0.003    0.998
```

```
## ParkAge:SurveyAngle     -13.84     4951.84  -0.003     0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.8558e+01  on 43  degrees of freedom
## Residual deviance: 1.5921e-08  on 40  degrees of freedom
## AIC: 8
##
## Number of Fisher Scoring iterations: 25
```

```r
logitAgeAngle2 <- glm(CompWinBinary ~ ParkAge  + SurveyAngle, data = bears, family = binomial)
summary(logitAgeAngle2)
```

```
##
## Call:
## glm(formula = CompWinBinary ~ ParkAge + SurveyAngle, family = binomial,
##     data = bears)
##
## Deviance Residuals:
##     Min        1Q     Median        3Q        Max
## -1.87409  -0.17119  -0.07917  -0.00294   1.15548
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -17.7677     7.7631  -2.289   0.0221 *
## ParkAge       0.6069     0.2783   2.181   0.0292 *
## SurveyAngle   0.8692     0.4897   1.775   0.0759 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.558  on 43  degrees of freedom
## Residual deviance: 19.207  on 41  degrees of freedom
## AIC: 25.207
##
## Number of Fisher Scoring iterations: 8
```

Idea for second hypothesis: filter for all who won first round. run logitmod with win and number of votes with filtered data

References

[1] Sarkar, D. (2021, October 08). A Glimpse Inside Fat Bear Week. Retrieved November 22, 2021, from https://www.discovermagazine.com/planet-earth/a-glimpse-inside-fat-bear-week

[2] Fat Bear Week 2021. (n.d.). Retrieved November 22, 2021, from https://explore.org/fat-bear-week