AMS 572 Project Report

Matthew Brulhardt, Anatoly Vitold Stankyavichyus,
Shenyi Ye, Ruijun Wu

November 17, 2018

Abstract

In this project, we mainly deal with an interesting candy data containing 85 samples and 12 variables with R. After doing data analysis and graphics, we formulate three meaningful hypotheses based on the results of data interpretation. These hypotheses are answered by *proportion test, t-test and logistic linear regression* and we analyze the results to get conclusions by different methods.

1 Summary of the Data

We choose this dataset from FiveThirtyEight website and you can view the dataset file *candy-data.csv* in github. "What's the best (or at least the most popular) Halloween candy?" That was the question this dataset was collected to answer. Data was collected by creating a website where participants were shown presenting two fun-sized candies and asked to click on the one they would prefer to receive. In total, more than 269 thousand votes were collected from 8,371 different IP addresses.

candy-data.csv includes attributes for each candy along with its ranking. For binary variables, 1 means yes, 0 means no. The data contains the following fields:

Header	***Description***
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutalmondy	Does it contain peanuts, peanut butter or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice, wafers, or a cookie component?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bag or box?
sugarpercent	The percentile of sugar it falls under within the data set.
pricepercent	The unit price percentile compared to the rest of the set.
winpercent	The overall win percentage according to 269,000 matchups.

From the raw data we can hardly see any relationships between so many features. So we need to analyze the data to help us understanding the variable correlation better and formulate meaningful hypotheses.

2 Data Interpretation

In this section, we mainly use R to analyze the data and plot correlation graphs of different features.

First, we have an overview of the data.

Figure 1: summary of our dataset

```
> summary(candy_data)
 competitorname      chocolate         fruity          caramel       peanutyalmondy       nougat
 Length:85          Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 Class :character   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
 Mode  :character   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000
                    Mean   :0.4353   Mean   :0.4471   Mean   :0.1647   Mean   :0.1647   Mean   :0.08235
                    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
                    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
 crispedricewafer      hard             bar            pluribus       sugarpercent     pricepercent
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0110   Min.   :0.0110
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.2200   1st Qu.:0.2550
 Median :0.00000   Median :0.0000   Median :0.0000   Median :1.0000   Median :0.4650   Median :0.4650
 Mean   :0.08235   Mean   :0.1765   Mean   :0.2471   Mean   :0.5176   Mean   :0.4786   Mean   :0.4689
 3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.7320   3rd Qu.:0.6510
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :0.9880   Max.   :0.9760
   winpercent
 Min.   :22.45
 1st Qu.:39.14
 Median :47.83
 Mean   :50.32
 3rd Qu.:59.86
 Max.   :84.18
```

From the outputs in figure 1, we can conclude that there seems to have no missing values in our dataset. Most candies are pluribus, meaning they are one of the many candies in the bag or box. More than two in every five candies either contains chocolate or has a fruity flavor. In order to see proportion of candies with different features more directly, we use the *ggplot* package in R to do data visualization.

From figure 2, we are able to see the differences among candies' features and conclude some facts:

(1).half of the candy in this dataset actually comes as many (pluribus feature);
(2).peanutyalmondy and nougat are not very popular features;
(3).there is a good balance of chocolate candies and fruity candies.

And then we have a peek at the rankings of each candy based on their winning percentage in figure 2. It seems like Reese's has dominated our top 10 candies. Another interesting finding on this ranking is that all candies in the top 10 contains chocolate and there is only one candy in our top 20 that doesn't contain chocolate (Starburst which has a fruity taste). Thus, we believe chocolate and fruity taste is mutually exclusive, they just don't go that well together.

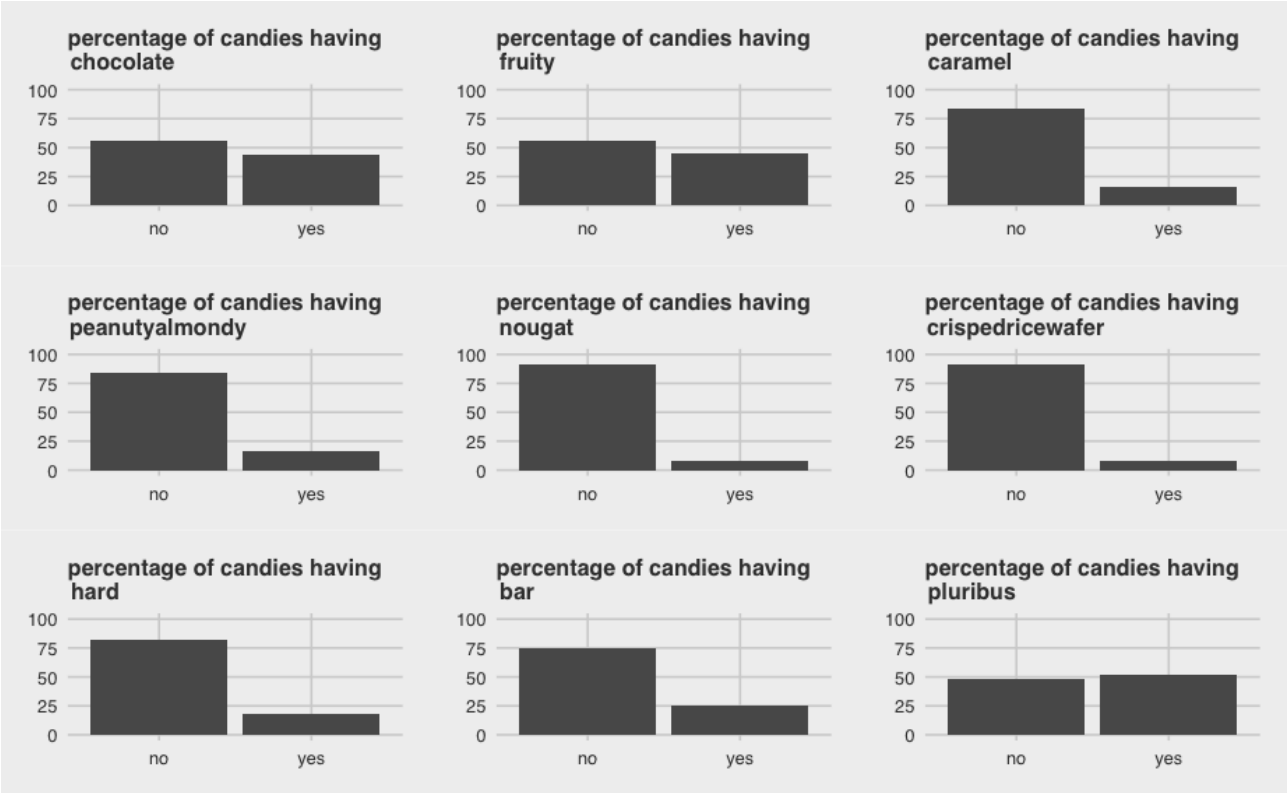


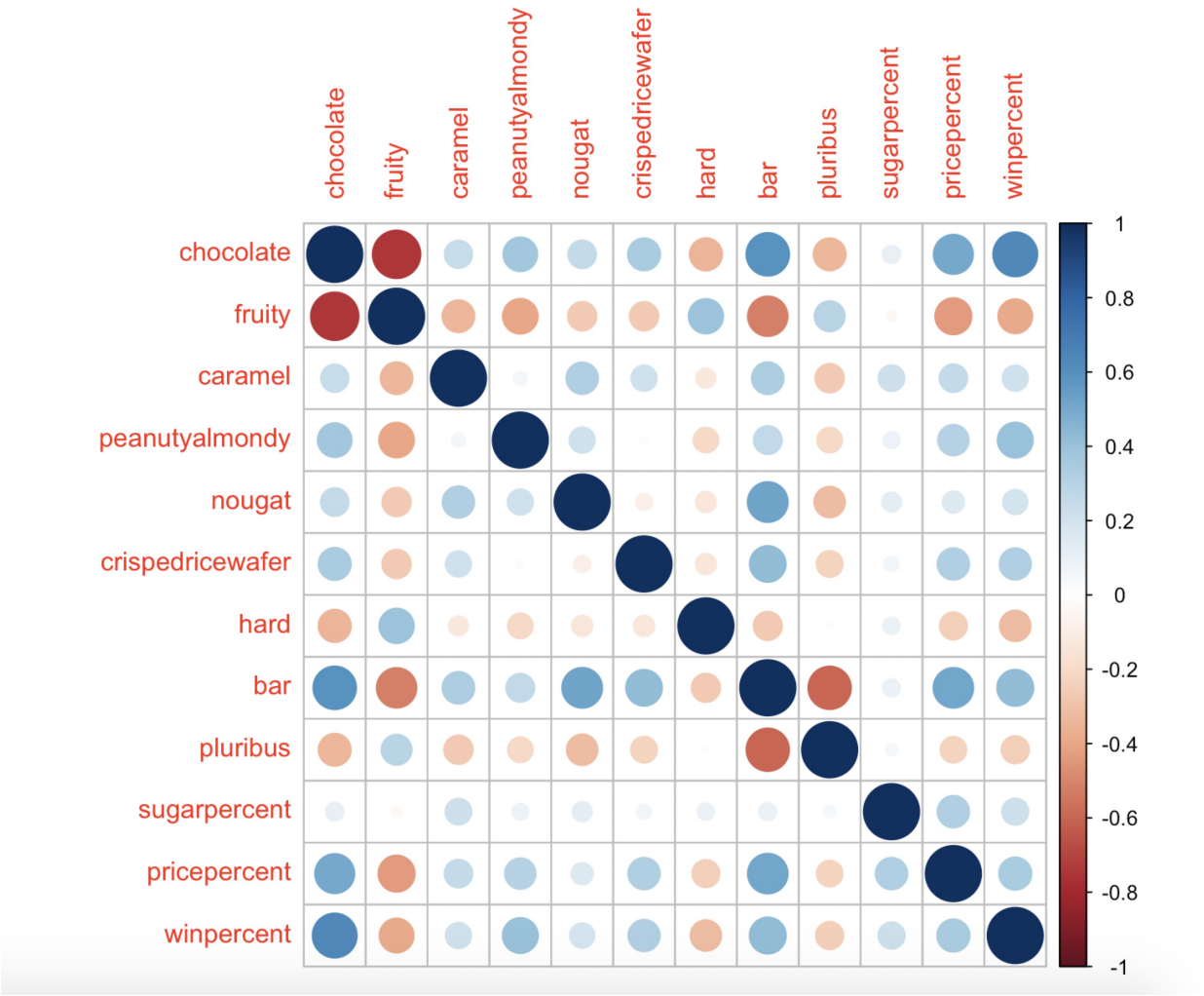
Figure 2: Percentage of Candies With Different Features

```
> candyrank[order(-winpercent),]
      competitorname                winpercent
 [1,] "ReeseÕs Peanut Butter cup"   "84.18029"
 [2,] "ReeseÕs Miniatures"          "81.866257"
 [3,] "Twix"                        "81.642914"
 [4,] "Kit Kat"                     "76.7686"
 [5,] "Snickers"                    "76.673782"
 [6,] "ReeseÕs pieces"              "73.43499"
 [7,] "Milky Way"                   "73.099556"
 [8,] "ReeseÕs stuffed with pieces" "72.887901"
 [9,] "Peanut butter M&MÕs"         "71.46505"
[10,] "Nestle Butterfinger"         "70.735641"
[11,] "Peanut M&Ms"                 "69.483788"
[12,] "3 Musketeers"                "67.602936"
[13,] "Starburst"                   "67.037628"
[14,] "100 Grand"                   "66.971725"
[15,] "M&MÕs"                       "66.574585"
[16,] "Nestle Crunch"               "66.47068"
[17,] "Rolo"                        "65.716286"
[18,] "Milky Way Simply Caramel"    "64.35334"
[19,] "Skittles original"           "63.08514"
[20,] "HersheyÕs Krackel"           "62.284481"
```

Figure 3: The 20 Most Popular Candy

In order to verify our observation, we make a correlation heatmap among our variables. The figure 4 is the heatmap in terms of correlation.

Figure 4: Correlation Heatmap



The figure above shows some interesting insights. Chocolate, Peanutyalmondy, Crispedrice-wafer, and Bar has the highest correlation among our attributes with winning percentage. However, Chocolate and fruity are most irrelevant features. It seems like the best chocolate would be a twix with peanuts or almonds instead of caramel.

Thus, based on what we get from the above, the hypotheses can be formulated more reasonable and meaningful.

3 Hypotheses

In this section, we are going to introduce our three hypotheses and draw the conclusion.

3.1 Hypothesis 1

Based on the figure 2, we know that percentages of candies having chocolate and fruity are higher than ones with other features and they are very similar. Also, they have a negative correlation which is close to -1. So, we want to check among the candies with winpercent greater than 50, if the proportion of chocolate is greater than fruity.

We set H_0 as null hypothesis and H_1 as alternative hypothesis:
H_0: $p_{chocolate} = p_{fruity}$ vs. H_1: $p_{chocolate} > p_{fruity}$

Then we do the proportional test and choose 95% confidence interval. Then the output using R is shown in figure 5.

```
> prop.test(x = c(xc, xf), n = c(nc, nf), alternative = "greater", conf.level = 0.95)

        2-sample test for equality of proportions with continuity correction

data:  c(xc, xf) out of c(nc, nf)
X-squared = 13.128, df = 1, p-value = 0.0001454
alternative hypothesis: greater
95 percent confidence interval:
 0.2426384 1.0000000
sample estimates:
   prop 1    prop 2
0.7179487 0.2820513
```

Figure 5: The output of proportional test

The test gives a p-value = 0.0001454, which is extremely small. So we reject H_0 and conclude that candies with chocolate has larger proportion than fruity candies among the "popular" candies.

3.2 Hypothesis 2

Based on the experience, sugar percentile of candies is decided by some ingredients, such as chocolate and caramel. So, we are interested in comparing the sugar among between candies with different ingredients. We would like to check if caramel candies contains more sugar than the chocolate candies. We make box plots for sugarpercent of caramel candies and chocolate candies.

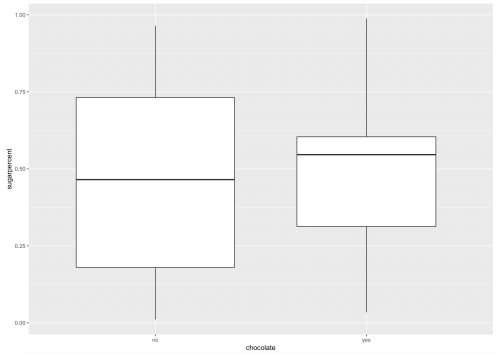


Figure 6: sugarpercent of chocolate and non-chocolate candies

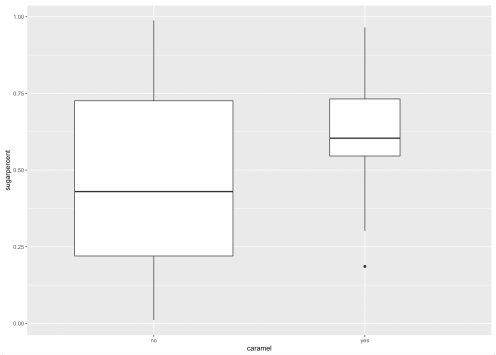


Figure 7: sugarpercent of caramel and non-caramel candies

From box plots we know that sugarpercent of caramel candies seems to be greater than chocolate candies. Thus, we set H_0 as null hypothesis and H_1 as alternative hypothesis to test it:

H_0: $\mu_{chocolate} = \mu_{caramel}$ vs. H_1: $\mu_{chocolate} < \mu_{caramel}$

We first check the normality of the two data using Shapiro method and perform a one-tailed t-test. The result of tests is shown below:

```
> #Test normality
> shapiro.test(chocolate_sugar)

        Shapiro-Wilk normality test

data:  chocolate_sugar
W = 0.9739, p-value = 0.524

> shapiro.test(caramel_sugar)

        Shapiro-Wilk normality test

data:  caramel_sugar
W = 0.95158, p-value = 0.5856

>
> #Perform a t-test:
> t.test(chocolate_sugar, caramel_sugar, alternative = 'less', mu=0, conf.level = 0.95)

        Welch Two Sample t-test

data:  chocolate_sugar and caramel_sugar
t = -1.483, df = 24.997, p-value = 0.07528
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.01626275
sample estimates:
mean of x mean of y
0.5120000 0.6191428
```

Figure 8: The output of equal mean t-test

From the output, we can conclude that sugarpercent of caramel and chocolate candies follows normal contribution at $\alpha = 0.05$. Since the p-value of two sample t-test is equal to 0.07528 which is greater than α. We should not reject H_0 and conclude that chocolate candies do not contain less sugar than the caramel candies.

3.3 Hypothesis 3

It seems that the correlation between chocolate and winpercent, chocolate and bar is relatively high. So we want know whether a candy is a bar, whether a candy is fruity and overall win percentage are good predictors of whether the candy contains chocolate or not. Because of the value of candies having chocolate, fruity and bar or not is binary, we choose *generalizedlinearmodel* to predict if these there variable can predict if a candy has chocolate or not. Then the result of logistic regression is shown below:

```
> summary(chocolate_model)

Call:
glm(formula = chocolate ~ bar + fruity + winpercent, family = "binomial",
    data = candy_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95672  -0.22792  -0.05484   0.15067   2.85234

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.11029    2.43102  -2.925 0.003447 **
bar1         0.99491    1.32275   0.752 0.451957
fruity1     -5.26724    1.55078  -3.397 0.000682 ***
winpercent   0.17000    0.05777   2.943 0.003254 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 116.407  on 84  degrees of freedom
Residual deviance:  31.504  on 81  degrees of freedom
AIC: 39.504

Number of Fisher Scoring iterations: 7
```

Figure 9: The output of equal mean t-test

To assess explained variation we will use McFadden's pseudo-R squared and construct GLE null mode. We get McFadden's pseudo-R squared is equal to 0.7293682 which is rather high. Therefore, it is reasonable to conclude that the information about whether a candy is a bar, whether a candy is fruity and overall win percentage is a good predictors of whether the candy contains chocolate or not.

4 Conclusion

From what we've done above, we can know if a candy has chocolate, it must be a popular Halloween candy among children. So if a producer wants their candies more popular, they should add some chocolate in it. Also a candy with chocolate usually does not contain less sugar than candy with caramel. Parents should notice about that. With the last hypothesis, we can predict whether the candy has chocolate or not by using some features. It will also a useful result for manufactures' decision.

In this project, we have learned how to work as a group, how to broke the project down into several parts in order to make this project more efficient and everyone can pool their knowledge and skills together. What's more, this project strengthens our understanding of what data analysis is, how to convert what we've learned at class to actual application , ability of using R to solve our hypothesis, using LaTeX to write our report. To analyze a data, we can choose different variables and methods. But what is the most important part is make your hypotheses reasonable and meaningful.

5 R Code

```r
library(ggplot2)
library(dplyr)
library(gridExtra)
library(viridis)
library(ggthemes)
library(ggridges)
library(RColorBrewer)
library(ggrepel)
library(ggdendro)
library(grid)
library(gridExtra)


candy_data<-read.csv("candy-data.csv",sep=",",stringsAsFactors=F)

summary(candy_data)

#------------------ggplot-barplot of percentage of candies with features----------

candy_data_1<-candy_data %>% dplyr::mutate(sum = rowSums(.[2:10]))

#convert to factor,1 means yes, 0 means no.
for(i in 2:10){
  candy_data_1[,i]<-ifelse(candy_data_1[,i]==1,'yes','no')
}
features<-c('chocolate', 'fruity', 'caramel', 'peanutyalmondy', 'nougat', 'crispedricewafer', 'hard', 'bar', 'pluribus')

# made the 3*3 plot
histPlot<-list()
for(i in 1:length(features)){
  histPlot[[i]]<-candy_data_1 %>% dplyr::group_by_(.dots =features[i]) %>%
    summarize(count=n()) %>%
    mutate(perc = round(100*count/sum(count),2)) %>%
    ggplot(aes_string(x=features[i])) +
    geom_bar(aes(y=perc), stat='identity') +
    theme_fivethirtyeight() +
    ggtitle(paste0('percentage of candies having\n',features[i])) +
    theme(plot.title=element_text(face="bold",hjust=.012,vjust=.8,colour="#3C3C3C",size=12)) + ylim(0,100)
}
do.call(grid.arrange, c(histPlot, ncol=3))

#---------------------winpercent ranking ---------
attach(candy_data)
candyrank<-cbind(competitorname,winpercent)
candyrank[order(-winpercent),]
```

```r

#-------------------correlation heat map-------------

library(corrplot) #Visualization of correlation
library(relaimpo) #Relative Importance of Attributes (Shapley Value)


#Correlation Heatmap of our data
candydatacor<-cor(candy_data[,-1])
corrplot(candydatacor)


#--------------------hypothesis 2   chocolate vs. caramel box polt--------------------
features <- candy_data_1 %>% select(2:10)
features[] <- lapply(features, as.logical)
ggplot(features, aes(x = chocolate, y= sugarpercent ) ) + geom_boxplot(varwidth = TRUE) # add the barpot
ggplot(features, aes(x = caramel, y= sugarpercent ) ) + geom_boxplot(varwidth = TRUE) # add the barpot
```

```r
candy_data$chocolate <- factor(candy_data$chocolate)
candy_data$fruity <- factor(candy_data$fruity)
candy_data$bar <- factor(candy_data$bar)

# Construct GLE model
chocolate_model <- glm(chocolate ~ bar + fruity + winpercent,
                       data = candy_data, family = "binomial")
summary(chocolate_model)


# To assess explained variation we will use McFadden's pseudo-R squared.

# Construct GLE null model
chocolate_predictor_null <- glm(chocolate ~ 1, data = candy_data, family="binomial")

#  McFadden's pseudo-R squared is equal to 0.7293682
1 - logLik(chocolate_model) / logLik(chocolate_predictor_null)

# Since pseudo-R squared is rather high, it is reasonable to conclude that the information
# about whether a candy is a bar, whether a candy is a fruity, and overall win percentage
# is a good predictors of whether the candy contains chocolate or not.
```

```r
#   Import data for the csv file
setwd("~/Desktop/AMS 572 project")
candy_data <- read.table("candy-data.csv", header=TRUE, sep=",")

#_____________________________________Data analysis and Plots_____________________________________________



#-------------------------Proportional Test Added-------11/02/2018---------------------------------

#Hypothesis 1

#Among the candies with winpercent greater than 50, the proportion of chocolate is greater than fruity

#H0: p.chocolate = p.fruity
#H1: p.chocolate > p.fruity

#First we extract data with winpercent greater than 50:
high_win = candy_data[which(candy_data$winpercent>50),]

#Then we collect the chocolate and fruity columns:
high_win_chocolate = high_win$chocolate
high_win_fruity = high_win$fruity

#Set up the parameters needed for the proportion test:
#Sample size of chocolate (n1):
nc = length(high_win_chocolate)
#Sample size of fruity (n2):
nf = length(high_win_fruity)
#Size of chocolate = 1:
xc = length(high_win_chocolate[which(high_win_chocolate==1)])
#Size of fruity = 1:
xf = length(high_win_fruity[which(high_win_fruity==1)])
#Proportion of chocolate (p1):
p.chocolate = xc/nc
#Proportion of fruity (p2):
p.fruity = xf/nf

#Do the proportion test with alpha = 0.05:
prop.test(x = c(xc, xf), n = c(nc, nf), alternative = "greater", conf.level = 0.95)

#The test gives a p.value = 0.0001454, which is extremely small, so we reject H0 and conclude that
#candies with chocolate has larger proportion than fruity candies among the "popular" candies.
```

```r


#_____________________________________________________________________________


#-------------------------------Equal Mean Test Added------11/08/2018-----------------------------------

#Hypothesis 2

#We are interested in comparing the sugar amount between candies with different ingredients.
#We would like to check if chocolate candies contains less sugar than the caramel candies.

#H0: mu.chocolate = mu.caramel
#H1: mu.chocolate < mu.caramel

#First we extract the data of chocolate and fruity candies into two groups:
chocolate_test2 = candy_data[which(candy_data$chocolate == 1),]
caramel_test2 = candy_data[which(candy_data$caramel == 1),]

#Then we get the corresponding sugar percentages:
chocolate_sugar = chocolate_test2$sugarpercent
caramel_sugar = caramel_test2$sugarpercent

#Test normality
shapiro.test(chocolate_sugar)
shapiro.test(caramel_sugar)

#Perform a t-test:
t.test(chocolate_sugar, caramel_sugar, alternative = 'less', mu=0, conf.level = 0.95)
#The test gives a p-value = 0.9247, so we do not reject H0 and conclude that there is not enough evidence
#which shows that chocolate candies contain more sugar than caramel candies.



#_____________________________________________________________________________

# Hypothesis 3:

# Whether a candy is a bar, whether a candy is fruity and overall win percentage are good predictors of whether the candy contains chocolate or not.

# Let's apply Generalized Linear Model to analyse this question.

# First, let's convert binary variables to factors
candy_data$chocolate <- factor(candy_data$chocolate)
```