Class10: Halloween Candy Mini Project

Brittney Hayes

2024-02-12

# 1. Importing Candy Data

candy\_data <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/candy-data.csv")   
candy\_file <- candy\_data

Q1. How many different candy types are in this dataset?

dim(candy\_file)

## [1] 85 13

ncol(candy\_file[2:10])

## [1] 9

There are 85 brands of candy and 9 types, I omitted a few columns that were not specifying types of candy.

Q2. How many fruity candy types are in the dataset?

sum(candy\_file["fruity"])

## [1] 38

There are 38 fruity candy types.

# 2. What is your favorite candy?

Q3. What is your favorite candy in the dataset and what is it’s winpercent value?

candy\_file[65,]$winpercent

## [1] 76.67378

My favorite candy is Snickers and the winpercent value is 76.67378.

Q4. What is the winpercent value for “Kit Kat”?

candy\_file[29,]$winpercent

## [1] 76.7686

Q5. What is the winpercent value for “Tootsie Roll Snack Bars”?

candy\_file[78,]$winpercent

## [1] 49.6535

#install.packages("skimr")  
library("skimr")  
skim(candy\_file)

Data summary

|  |  |
| --- | --- |
| Name | candy\_file |
| Number of rows | 85 |
| Number of columns | 13 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| competitorname | 0 | 1 | 4 | 27 | 0 | 85 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| chocolate | 0 | 1 | 0.44 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▆ |
| fruity | 0 | 1 | 0.45 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▆ |
| caramel | 0 | 1 | 0.16 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| peanutyalmondy | 0 | 1 | 0.16 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| nougat | 0 | 1 | 0.08 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| crispedricewafer | 0 | 1 | 0.08 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| hard | 0 | 1 | 0.18 | 0.38 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| bar | 0 | 1 | 0.25 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| pluribus | 0 | 1 | 0.52 | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | ▇▁▁▁▇ |
| sugarpercent | 0 | 1 | 0.48 | 0.28 | 0.01 | 0.22 | 0.47 | 0.73 | 0.99 | ▇▇▇▇▆ |
| pricepercent | 0 | 1 | 0.47 | 0.29 | 0.01 | 0.26 | 0.47 | 0.65 | 0.98 | ▇▇▇▇▆ |
| winpercent | 0 | 1 | 50.32 | 14.71 | 22.45 | 39.14 | 47.83 | 59.86 | 84.18 | ▃▇▆▅▂ |

Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

library(skimr)  
skim(candy\_file)

Data summary

|  |  |
| --- | --- |
| Name | candy\_file |
| Number of rows | 85 |
| Number of columns | 13 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| competitorname | 0 | 1 | 4 | 27 | 0 | 85 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| chocolate | 0 | 1 | 0.44 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▆ |
| fruity | 0 | 1 | 0.45 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | ▇▁▁▁▆ |
| caramel | 0 | 1 | 0.16 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| peanutyalmondy | 0 | 1 | 0.16 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| nougat | 0 | 1 | 0.08 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| crispedricewafer | 0 | 1 | 0.08 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▁ |
| hard | 0 | 1 | 0.18 | 0.38 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| bar | 0 | 1 | 0.25 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |
| pluribus | 0 | 1 | 0.52 | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | ▇▁▁▁▇ |
| sugarpercent | 0 | 1 | 0.48 | 0.28 | 0.01 | 0.22 | 0.47 | 0.73 | 0.99 | ▇▇▇▇▆ |
| pricepercent | 0 | 1 | 0.47 | 0.29 | 0.01 | 0.26 | 0.47 | 0.65 | 0.98 | ▇▇▇▇▆ |
| winpercent | 0 | 1 | 50.32 | 14.71 | 22.45 | 39.14 | 47.83 | 59.86 | 84.18 | ▃▇▆▅▂ |

Sugarpercent, pricepercent, and winpercent appear different.

Q7. What do you think a zero and one represent for the candy$chocolate column?

table(candy\_file$chocolate)

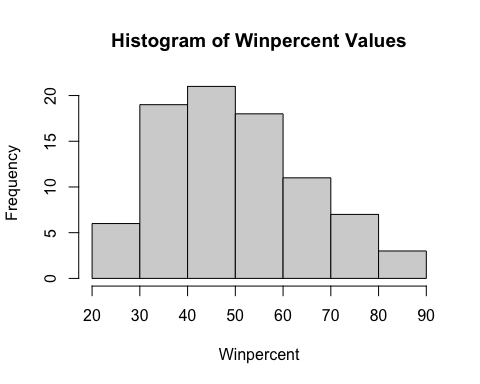
##   
## 0 1   
## 48 37

The 0 and 1 confirm if the candy type falls into this candy, 0 means it does not/false and 1 means it does/true. So 0 is not chocolate and 1 is chocolate.

# 3. Overall Candy Rankings

Q8. Plot a histogram of winpercent values

hist(candy\_data$winpercent,   
 main = "Histogram of Winpercent Values",  
 xlab = "Winpercent",  
 ylab = "Frequency")



Q9. Is the distribution of winpercent values symmetrical? Based on the appearance of the histogram, no.

Q10. Is the center of the distribution above or below 50%?

mean(candy\_file$winpercent)

## [1] 50.31676

median(candy\_file$winpercent)

## [1] 47.82975

Based on the mean, it’s above 50% but by very little. Based on the median, it’s below 50% again by little. Based on the histogram it appears below 50%. So I would say below 50%.

Q11. On average is chocolate candy higher or lower ranked than fruit candy?

candy\_file$winpercent[as.logical(candy\_file$chocolate)]

## [1] 66.97173 67.60294 50.34755 56.91455 38.97504 55.37545 62.28448 56.49050  
## [9] 59.23612 57.21925 76.76860 71.46505 66.57458 55.06407 73.09956 60.80070  
## [17] 64.35334 47.82975 54.52645 70.73564 66.47068 69.48379 81.86626 84.18029  
## [25] 73.43499 72.88790 65.71629 34.72200 37.88719 76.67378 59.52925 48.98265  
## [33] 43.06890 45.73675 49.65350 81.64291 49.52411

mean(candy\_file$winpercent[as.logical(candy\_file$chocolate)])

## [1] 60.92153

candy\_file$winpercent[as.logical(candy\_file$fruity)]

## [1] 52.34146 34.51768 36.01763 24.52499 42.27208 39.46056 43.08892 39.18550  
## [9] 46.78335 57.11974 51.41243 42.17877 28.12744 41.38956 39.14106 52.91139  
## [17] 46.41172 55.35405 22.44534 39.44680 41.26551 37.34852 35.29076 42.84914  
## [25] 63.08514 55.10370 45.99583 59.86400 52.82595 67.03763 34.57899 27.30386  
## [33] 54.86111 48.98265 47.17323 45.46628 39.01190 44.37552

mean(candy\_file$winpercent[as.logical(candy\_file$fruity)])

## [1] 44.11974

Chocolate candy is ranked higher than fruity candy on average.

Q12. Is this difference statistically significant?

chocolate <- candy\_file$winpercent[as.logical(candy\_file$chocolate)]  
fruit <- candy\_file$winpercent[as.logical(candy\_file$fruity)]  
  
t.test(chocolate,fruit)

##   
## Welch Two Sample t-test  
##   
## data: chocolate and fruit  
## t = 6.2582, df = 68.882, p-value = 2.871e-08  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 11.44563 22.15795  
## sample estimates:  
## mean of x mean of y   
## 60.92153 44.11974

Since the p-value is less than 0.05, yes the difference is statistically significant.

Q13. What are the five least liked candy types in this set?

head(candy\_file[order(candy\_file$winpercent),], n=5)

## competitorname chocolate fruity caramel peanutyalmondy nougat  
## 45 Nik L Nip 0 1 0 0 0  
## 8 Boston Baked Beans 0 0 0 1 0  
## 13 Chiclets 0 1 0 0 0  
## 73 Super Bubble 0 1 0 0 0  
## 27 Jawbusters 0 1 0 0 0  
## crispedricewafer hard bar pluribus sugarpercent pricepercent winpercent  
## 45 0 0 0 1 0.197 0.976 22.44534  
## 8 0 0 0 1 0.313 0.511 23.41782  
## 13 0 0 0 1 0.046 0.325 24.52499  
## 73 0 0 0 0 0.162 0.116 27.30386  
## 27 0 1 0 1 0.093 0.511 28.12744

The 5 least liked candy types are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

Q14. What are the top 5 all time favorite candy types out of this set?

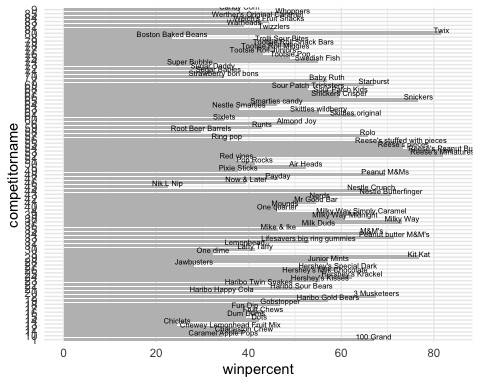
tail(candy\_file[order(candy\_file$winpercent),], n=5)

## competitorname chocolate fruity caramel peanutyalmondy nougat  
## 65 Snickers 1 0 1 1 1  
## 29 Kit Kat 1 0 0 0 0  
## 80 Twix 1 0 1 0 0  
## 52 Reese's Miniatures 1 0 0 1 0  
## 53 Reese's Peanut Butter cup 1 0 0 1 0  
## crispedricewafer hard bar pluribus sugarpercent pricepercent winpercent  
## 65 0 0 1 0 0.546 0.651 76.67378  
## 29 1 0 1 0 0.313 0.511 76.76860  
## 80 1 0 1 0 0.546 0.906 81.64291  
## 52 0 0 0 0 0.034 0.279 81.86626  
## 53 0 0 0 0 0.720 0.651 84.18029

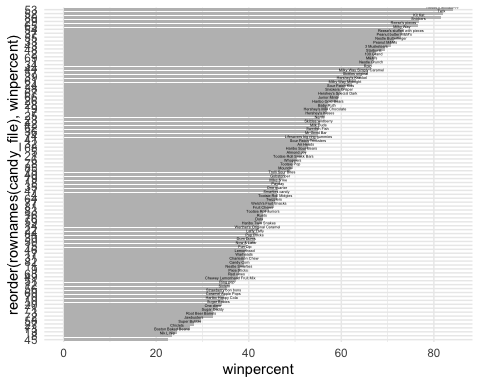
The top 5 all time favorite candy types are Reese’s Peanut Butter cup, Reese’s Miniatures, Twix, Kit Kat, and Snickers.

Q15. Make a first barplot of candy ranking based on winpercent values.

#install.packages("ggplot2")  
library(ggplot2)  
  
ggplot(candy\_file) +   
 aes(x = winpercent, y = rownames(candy\_file)) +  
 geom\_bar(stat = "identity", fill = "grey") +  
 labs(x = "winpercent", y = "competitorname") +  
 geom\_text(aes(label=competitorname), vjust=-0.1, size=2) +  
 theme\_minimal()

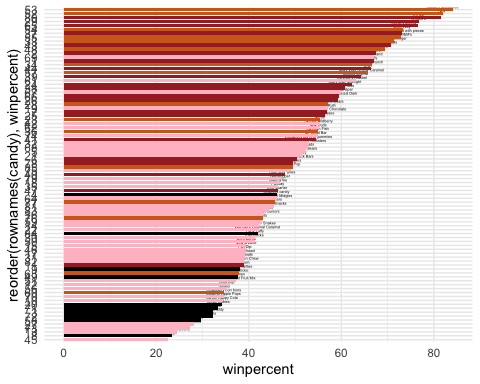
 Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

library(ggplot2)  
  
ggplot(candy\_file) +   
 aes(x = winpercent, y = rownames(candy\_file)) +  
 geom\_bar(stat = "identity", fill = "grey") +  
 labs(x = "winpercent", y = "competitorname") +  
 geom\_text(aes(label=competitorname), vjust=-2, size=1) +  
 theme\_minimal() + aes(winpercent, reorder(rownames(candy\_file),winpercent))



Sorted with color:

candy <- candy\_file  
my\_cols=rep("black", nrow(candy))  
my\_cols[as.logical(candy$chocolate)] = "chocolate"  
my\_cols[as.logical(candy$bar)] = "brown"  
my\_cols[as.logical(candy$fruity)] = "pink"  
  
ggplot(candy) +   
 aes(x = winpercent, y = rownames(candy)) +  
 geom\_bar(stat = "identity", fill = "grey") +  
 labs(x = "winpercent", y = "competitorname") +  
 geom\_text(aes(label=competitorname), vjust=-2, size=1) +  
 theme\_minimal() + aes(winpercent, reorder(rownames(candy),winpercent)) + geom\_col(fill=my\_cols)



Q17. What is the worst ranked chocolate candy? The worst ranked chocolate candy is Sixlets.

Q18. What is the best ranked fruity candy? The best ranked fruity candy is Starburst.