# Class 10: Halloween Candy Mini-Project

Eric Jordahl

2022-10-28

## Table of contents

## Background

In this mini project we will examine 538 halloween candy data.

```
url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/

candy <-  read.csv(url, row.names=1)
head(candy)
```

```
             chocolate fruity caramel peanutyalmondy nougat crispedricewafer
100 Grand            1      0       1              0      0                 1
3 Musketeers         1      0       0              0      1                 0
One dime             0      0       0              0      0                 0
One quarter          0      0       0              0      0                 0
Air Heads            0      1       0              0      0                 0
Almond Joy           1      0       0              1      0                 0
             hard bar pluribus sugarpercent pricepercent winpercent
100 Grand       0   1        0        0.732        0.860   66.97173
3 Musketeers    0   1        0        0.604        0.511   67.60294
One dime        0   0        0        0.011        0.116   32.26109
One quarter     0   0        0        0.011        0.511   46.11650
Air Heads       0   0        0        0.906        0.511   52.34146
Almond Joy      0   1        0        0.465        0.767   50.34755
```

## Q1.

**How many different candy types are in this dataset?**

```
nrow(candy)
```

```
[1] 85
```

There are 85 candies in this dataset

## Q2.

**How many fruity candy types are in the dataset?**

```
sum(candy$fruity)
```

```
[1] 38
```

There are 38 fruity candies in this dataset

**What is your favorite Candy?**

## Q3

**What is your favorite candy in the dataset and what is it's winpercent value?**

```
candy["Kit Kat", ]$winpercent
```

```
[1] 76.7686
```

My favorite candy is Kit Kat and the win percentage is 76.7686.

## Q4

**What is the winpercent value for "Kit Kat"?**

The win percent value for Kit Kat 76.7686.

## Q5

**What is the winpercent value for Tootsie Roll Snack Bars?**

The win percent value for Tootsie Roll Snack Bars is 49.653503.

```
library("skimr")
skim(candy)
```

Table 1: Data summary

| | |
|---|---|
| Name | candy |
| Number of rows | 85 |
| Number of columns | 12 |
| _____ | |
| Column type frequency: | |
| numeric | 12 |
| _____ | |
| Group variables | None |

**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?**

The win percentage average is much higher than most of the other columns as it is in percentage not fraction of the data that is chosen.

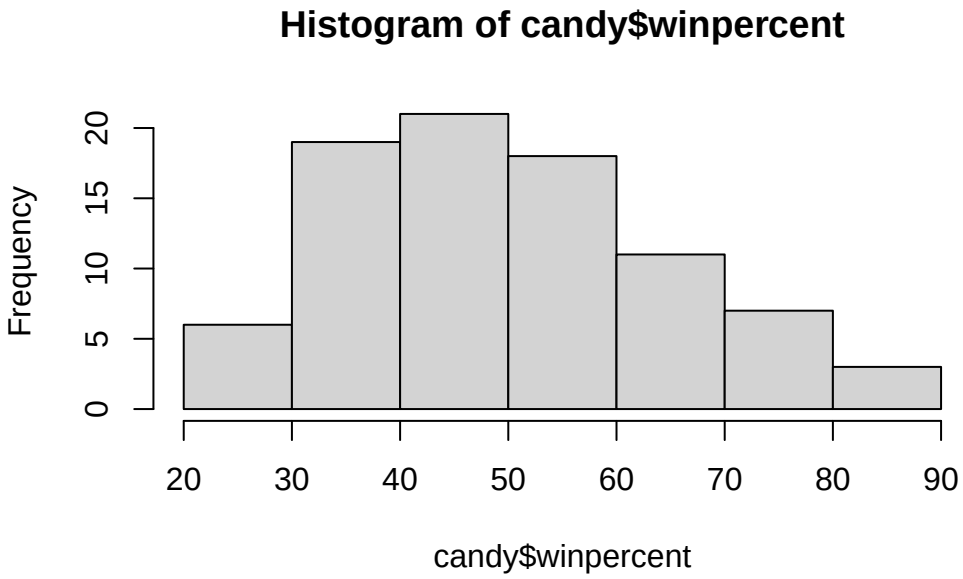
## Q7

**What do you think a zero and one represent for the candy$chocolate column?** The zero and one would represent either yes or no for if the candy is chocolate or not,

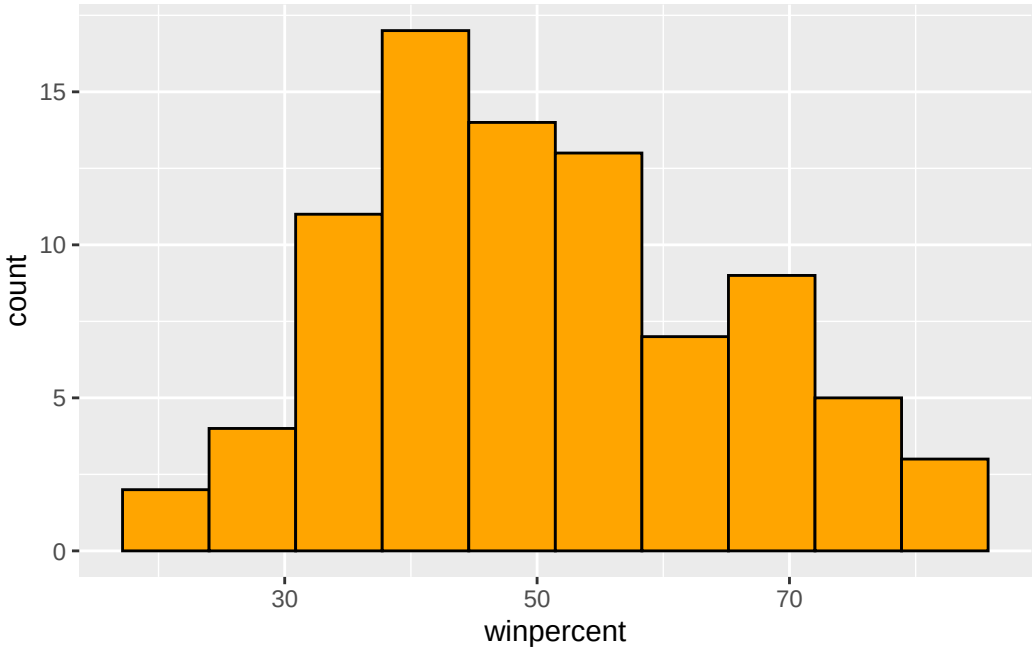
## Q8

**Plot a histogram of winpercent values**

```
hist(candy$winpercent)
```




```
library(ggplot2)
ggplot(candy, aes(winpercent)) + geom_histogram(bins=10, fill = "orange", col="black")
```




## Q9

**Is the distribution of winpercent values symmetrical?**

No, the distribution is not symmetrical

## Q10

**Is the center of the distribution above or below 50%?** The center of the distribution is below 50%, as the center/median is 47.829754

#**Q11 On average is the cholocate candy higher or lower ranked than fruit candy?**

```
choc_mean_winper <- mean(candy$winpercent[as.logical(candy$chocolate)])
fruit_mean_winper <- mean(candy$winpercent[as.logical(candy$fruity)])
choc_mean_winper > fruit_mean_winper
```

```
[1] TRUE
```

## Q12

**Is this difference statistically significant?**

```
  choc_stats <- candy$winpercent[as.logical(candy$chocolate)]
  fruit_stats <- candy$winpercent[as.logical(candy$fruity)]
  t.test(choc_stats, fruit_stats)
```

```
    Welch Two Sample t-test

data:  choc_stats and fruit_stats
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
```

Yes this is statistically significant becasue the p-value is less than .05

### Overall Candy Rankings

## Q13

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

```
  library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
candy %>% arrange(winpercent) %>% head(5)
```

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

The least liked candies 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?**

```
library (dplyr)

candy %>% arrange(winpercent) %>% tail(5)
```

```
                        chocolate fruity caramel peanutyalmondy nougat
Snickers                        1      0       1              1      1
Kit Kat                         1      0       0              0      0
Twix                            1      0       1              0      0
ReeseÕs Miniatures              1      0       0              1      0
ReeseÕs Peanut Butter cup       1      0       0              1      0
                        crispedricewafer hard bar pluribus sugarpercent
```

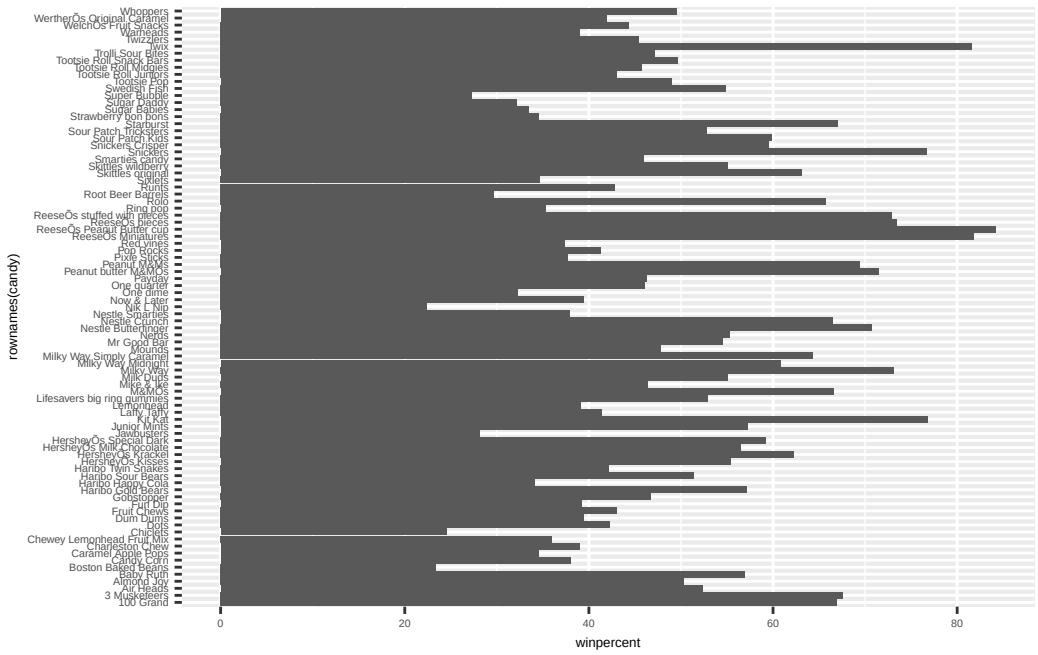
```
Snickers                            0    0    1        0        0.546
Kit Kat                             1    0    1        0        0.313
Twix                                1    0    1        0        0.546
ReeseÕs Miniatures                  0    0    0        0        0.034
ReeseÕs Peanut Butter cup           0    0    0        0        0.720
                          pricepercent winpercent
Snickers                         0.651   76.67378
Kit Kat                          0.511   76.76860
Twix                             0.906   81.64291
ReeseÕs Miniatures               0.279   81.86626
ReeseÕs Peanut Butter cup        0.651   84.18029
```

The top 5 candies are Snickers, Kit Kat, Twix, Reese's Miniatures, and Reese's Peanut Butter Cup

## Q15.

### Make a first barplot of candy ranking based on winpercent values

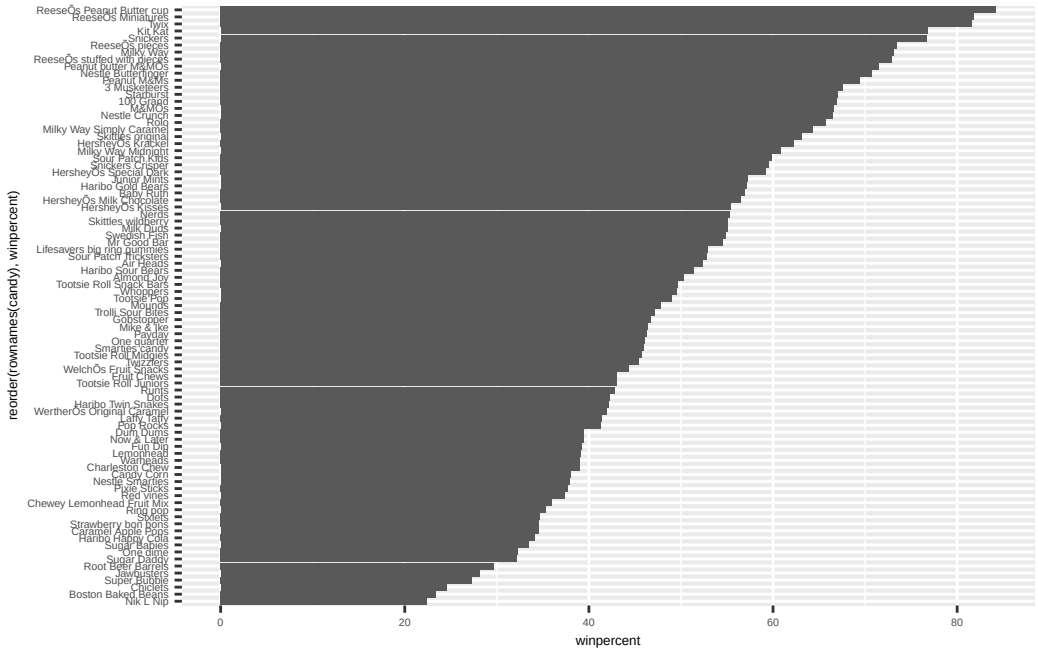
```
ggplot(candy) + aes(winpercent, rownames(candy)) + geom_col() + theme(text=element_text(si
```



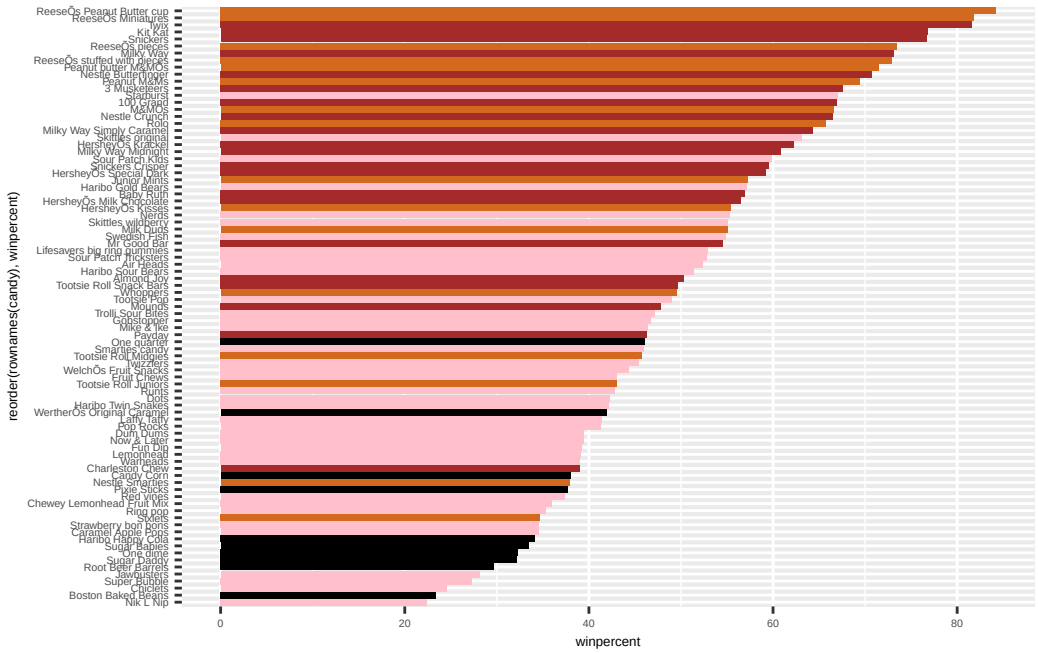
## Q16

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

```
ggplot(candy) + aes(winpercent, reorder(rownames(candy),winpercent)) + geom_col() + theme(
```




```
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(winpercent, reorder(rownames(candy),winpercent), ) + geom_col(fill=my_
```




## Q17.

**What is the worst ranked chocolate candy?**

The worst ranked chocolate candy is "Sixlets"

## Q18

**What is the best ranked fruity candy?**

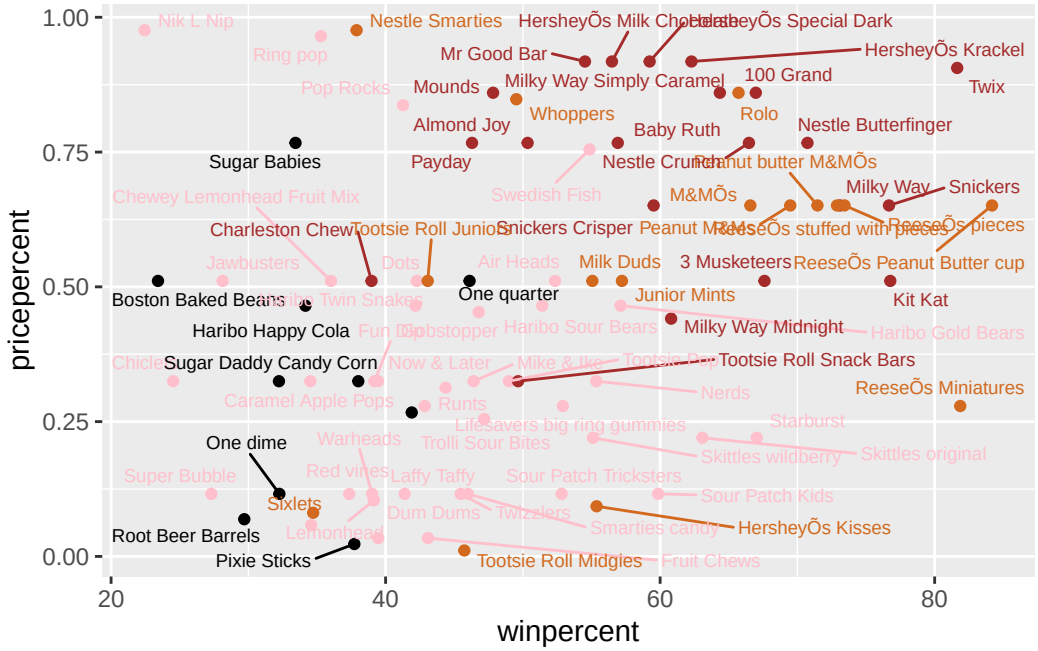
The Best ranked fruity candy is "Starburst"

### Taking a look at Pricepercent

```
library(ggrepel)

# How about a plot of price vs win
ggplot(candy) +
 aes(winpercent, pricepercent, label=rownames(candy)) +
```

```
    geom_point(col=my_cols) +
    geom_text_repel(col=my_cols, size=2.5, max.overlaps = 15)
```

```
Warning: ggrepel: 3 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




**#Q19 Which candy type is the highest ranked in terms of winpercent for the least money - i.e. offers the most bang for your buck?**

The candy that is ranked highest with the lowest price is the Reese's Miniatures as it has a high win percentage with an overall low price

## Q20

**What are the top 5 most expensive candy types in the dataset and of these which is the least popular?**

```
library (dplyr)

most_exp_candy <-candy %>% arrange(pricepercent) %>% tail(5)
```

```r
most_exp_candy_order <- most_exp_candy %>% arrange(winpercent) %>% head(1)
most_exp_candy_order
```

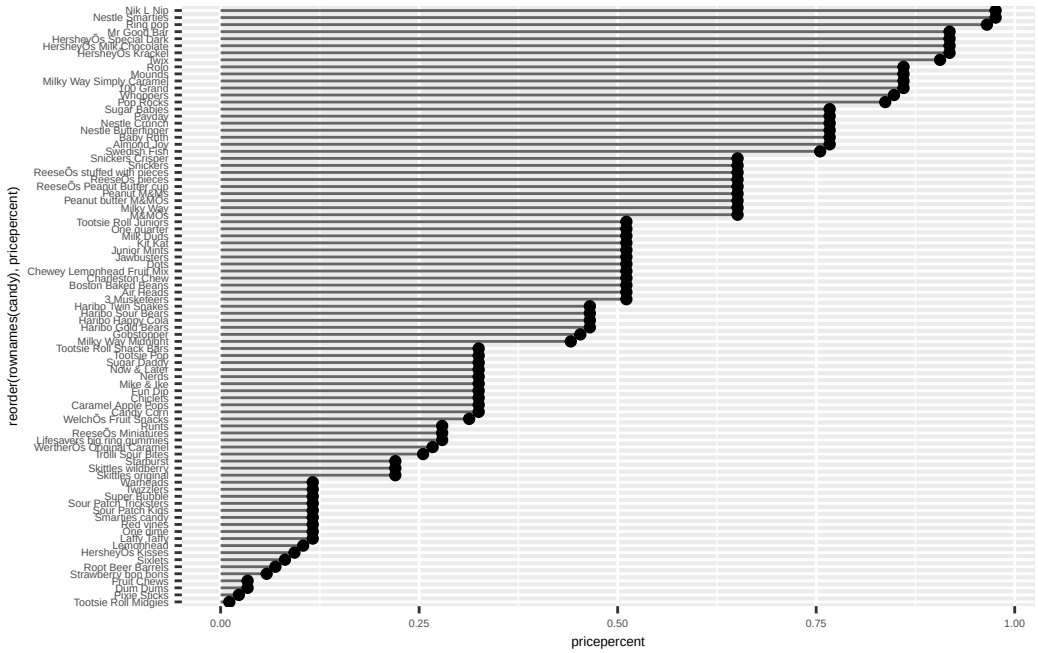
```
          chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
Nik L Nip         0      1       0              0      0                 0    0
          bar pluribus sugarpercent pricepercent winpercent
Nik L Nip   0        1        0.197        0.976   22.44534
```

Nik L Nip has the lowest ranking (by winpercent) of the 5 most expensive candies.

## Q21

### Make a lollipop chart of the data by pricepercent vs candy

```r
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point() + theme(text=element_text(size=5))
```

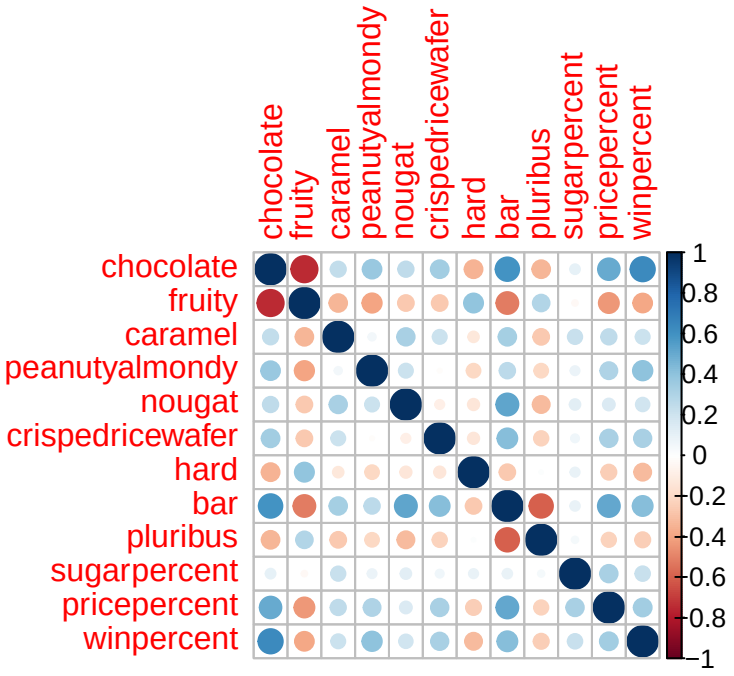


### Exploring the Correlation Structure

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
cij <- cor(candy)
corrplot(cij)
```




## Q22

**Examining this plot what two variables are anti-correlated (i.e. have minus values)?**

Chocolate and Fruity are the most negatively correlated of these categories

## Q23

**Similarly, what two variables are most positively correlated?**

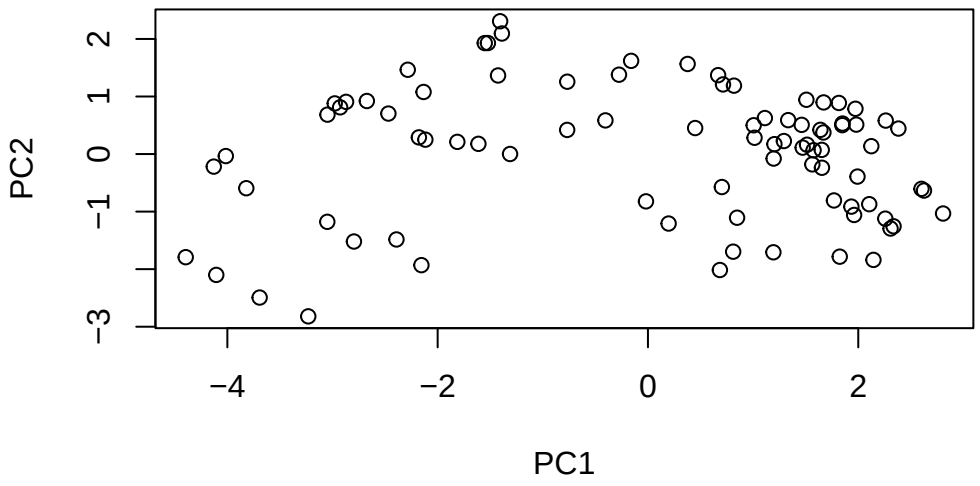
Chocolate and winpercentage or bar are very positively correlated

### Principle Component Analysis

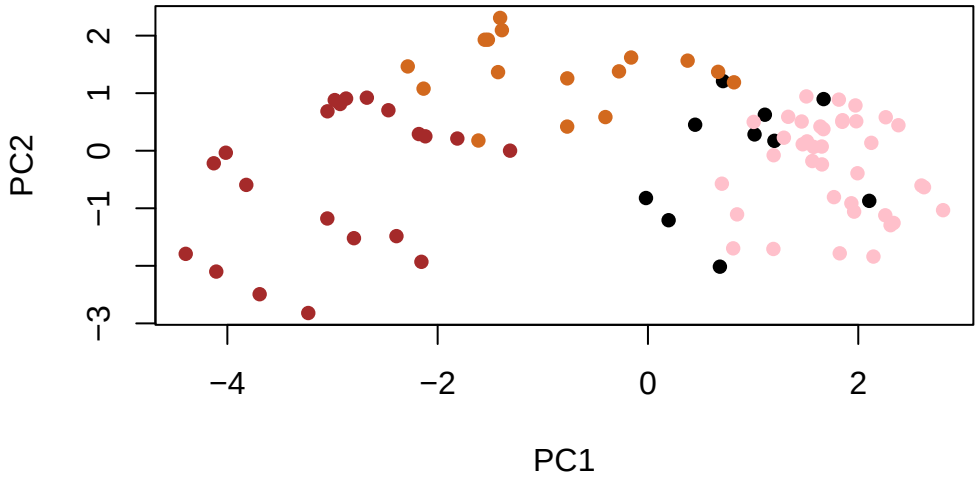
```
pca <- prcomp(candy, scale=T)
summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 0.7424 0.79830 0.85369
                           PC8     PC9    PC10    PC11    PC12
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

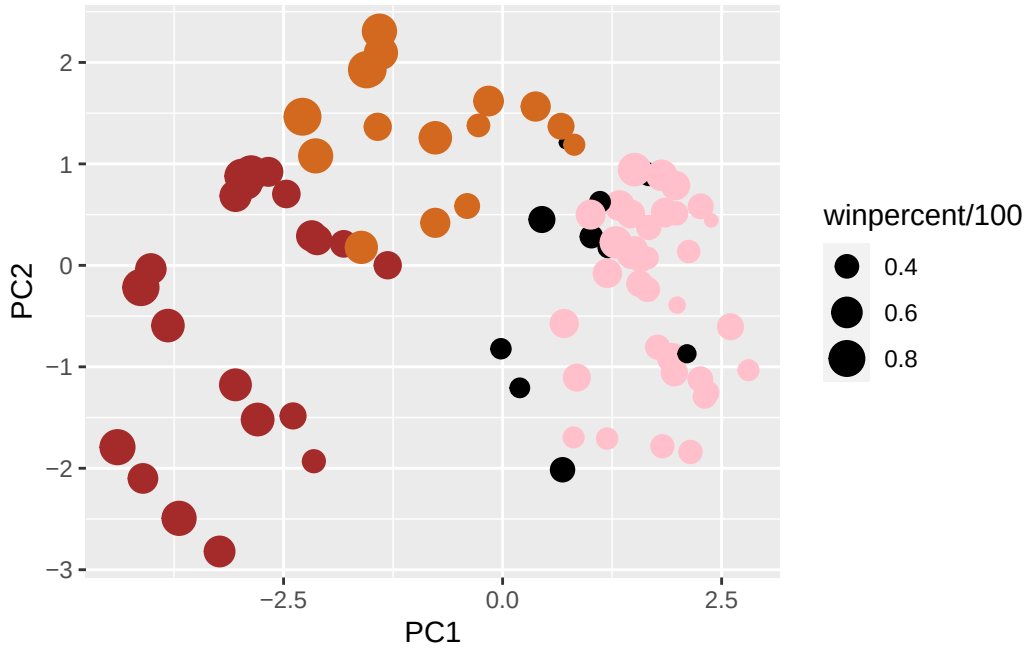
```
plot(pca$x[,1:2])
```



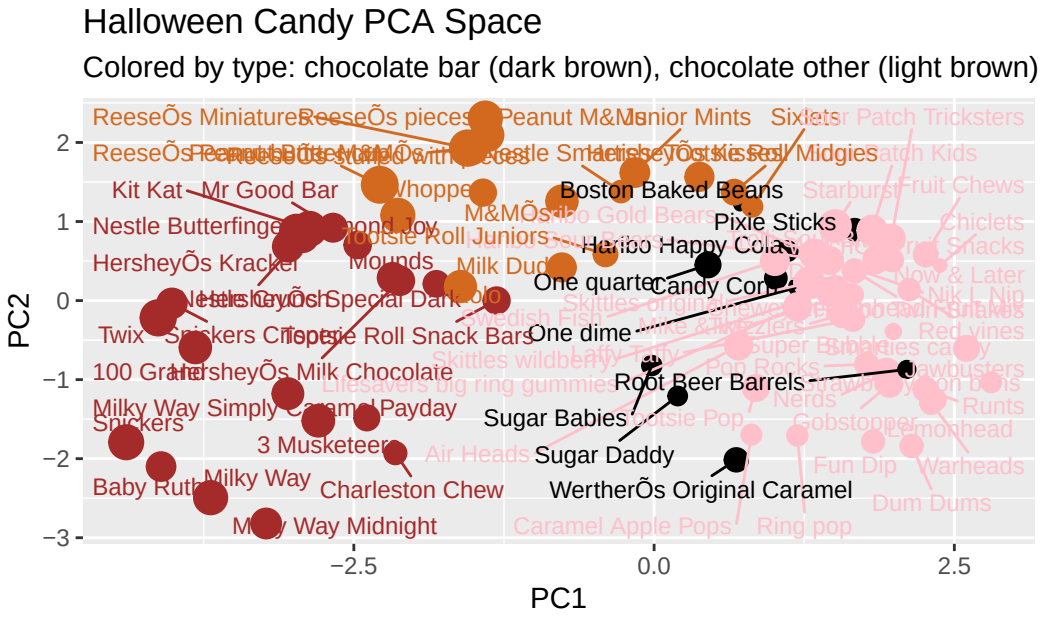

```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



```
my_data <- cbind(candy, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```




```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 100)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
      subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown)
       caption="Data from 538")
```




```
library(plotly)
```

```
Attaching package: 'plotly'
```

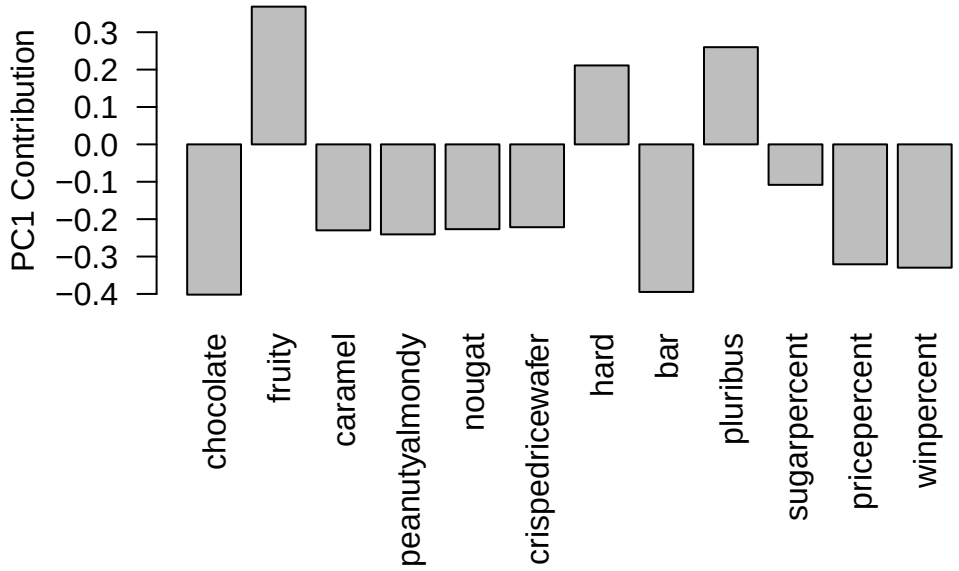
```
The following object is masked from 'package:ggplot2':

    last_plot
```

```
The following object is masked from 'package:stats':

    filter
```

```
The following object is masked from 'package:graphics':

    layout
```

```
#ggplotly(p)
```

```
par(mar=c(8,4,2,2))
barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```




## Q24

**What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?**

The original variables that are picked up are fruity, hard, and pluribus Yes I would say this generally makes sense as fruity candies tend to have many in a package and they tend to be harder than many other candies. They are shown as also positively correlated in our correlation plot.