Class 10 Lab: Halloween Mini-Project

Joshua Mac

2025-02-05

538 Halloween Candy Data Set!

Importing candy data

First, fetch the data:

```
candy_file <-"https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-rank
candy = read.csv(candy_file, 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
```

The dataset contains different kinds of candy sorted by many different characteristics such as whether it has chocolate, if it has a fruit flavor, its sugar content, etc.

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

A1. There are 85 different candy types in the dataset. See the code below.

```
nrow(candy)
```

```
[1] 85
```

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

A2. There are 38 fruity candy types in the dataset. See the code below.

```
sum(candy$fruity==1)
```

```
[1] 38
```

What is my favorite candy?

Exploring winpercent variable

```
candy["Twix", ]$winpercent
```

```
[1] 81.64291
```

This variable corresponds to the percentage of people who prefer the candy over another randomly chosen candy from the dataset, pretty neat!

Q3. What is your favorite candy in the dataset and what is its winpercent value?

A3. My favorite candy is Hershey's Kisses; its winpercent value is 55.37545. See the code below.

```
candy["Hershey's Kisses", ]$winpercent
```

```
[1] 55.37545
```

Q4/5. What is the winpercent value for "Kit Kat"? for "Tootsie Roll Snack Bars"?

A4/5. The winpercent values are 76.7686 and 49.6535, respectively.

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

```
[1] 76.7686
```

```
candy["Tootsie Roll Snack Bars", ]$winpercent
```

```
[1] 49.6535
```

Let's use skim() from the skimr package to give a quick overview of the dataset.

```
# install.packages("skimr")
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	

At a glance, the overview includes mean and sd, which saves us having to find them ourselves.

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

A6. Yes, winpercent seems to be on a completely different scale to the majority.

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

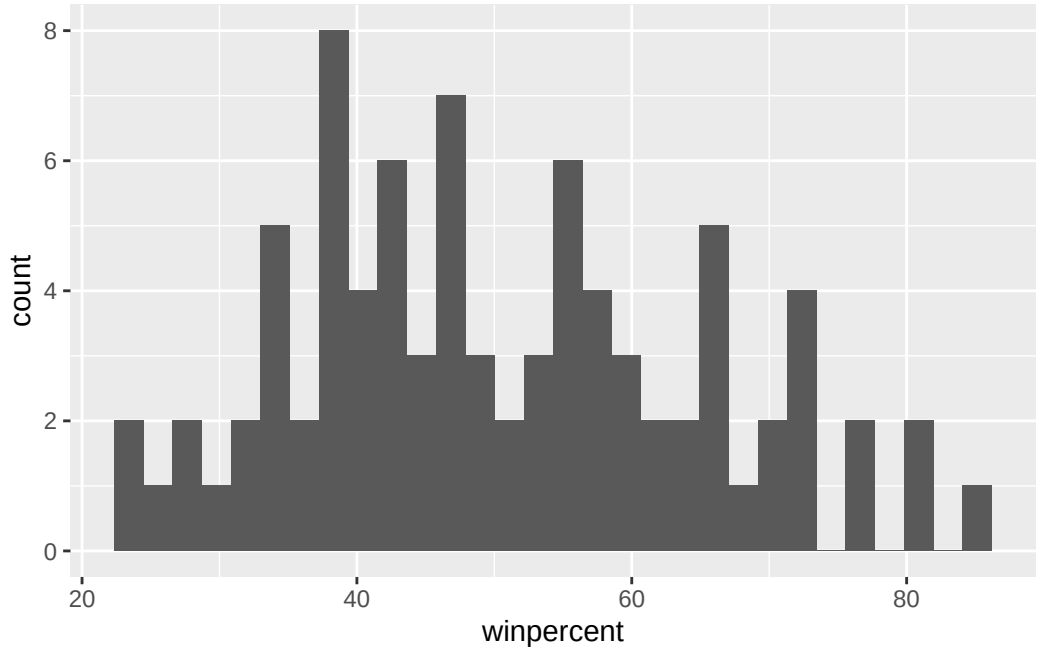
A7. The zero and one most likely represent "no" and "yes", respectively, as in binary. This can be understood through the meaning of the chocolate variable and evaluating a candy that is and is not chocolatey against each other. For example, 3 Musketeers has a value of 1 for the chocolate column because it contains chocolate while Air Heads, a fruity taffy candy, has a value of 0.

Q8. Plot a histogram of winpercent values

A8. See below code/histogram

```
library(ggplot2)
ggplot(candy, aes(winpercent))+
  geom_histogram()
```

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



Q9. Is the distribution of winperecent values symmetrical>

A9. The distribution of winpercent values is not symmetrical, there is a slight skew to the right.This means the mean will be greater than the median as the values at the higher end influence mean more than median changes with an addition of a new point in the same upper range.

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

```
median(candy$winpercent)
```

```
[1] 47.82975
```

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

A10. The center of the distribution (median) is below 50%, at approximately 47.83. See above actually^

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

A11. On average, chocolate candy (M=60.92153) is ranked higher than fruit candy (M=44.11974).

```
mean(candy$winpercent[as.logical(candy$chocolate)]) > mean(candy$winpercent[as.logical(candy$
```

```
[1] TRUE
```

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

Q12. Is this difference statistically significant?

A12. No, the difference is not statistically significant, as the p-value is > 0.05.

```
t.test(candy$chocolate, candy$fruity)
```

```
    Welch Two Sample t-test

data:  candy$chocolate and candy$fruity
t = -0.15357, df = 168, p-value = 0.8781
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1630081  0.1394786
sample estimates:
mean of x mean of y
0.4352941 0.4470588
```

Overall Candy Rankings

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

> A13. The five least liked are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters

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

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

A14. The top 5 are Snickers, Kit Kat, Twix, Reese's Minatures, and Reese's PBC

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

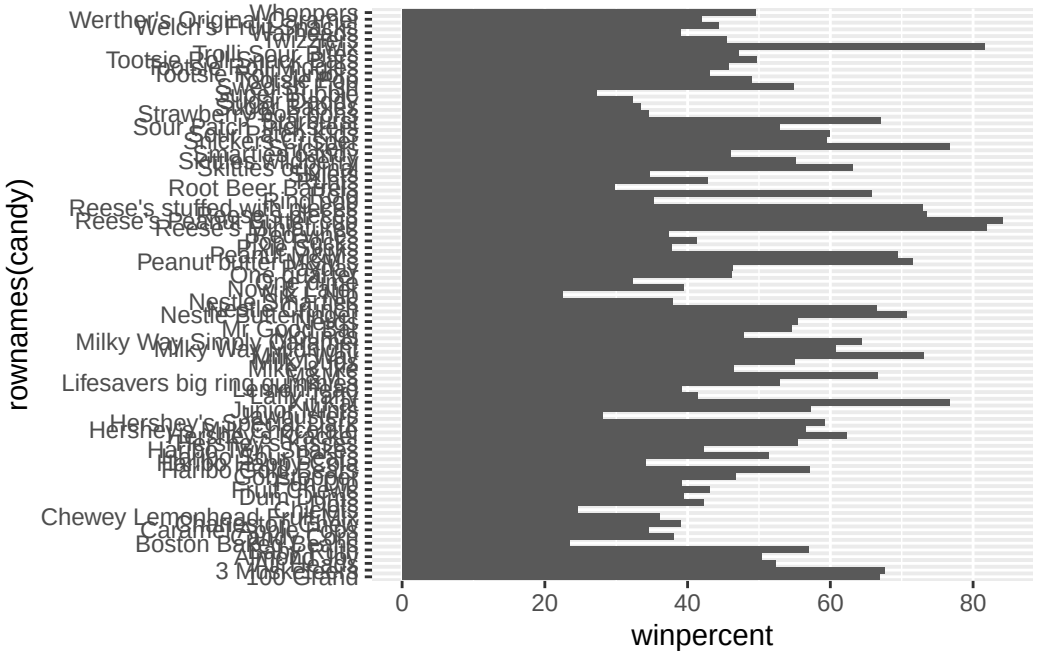
	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
Snickers	1	0	1	1	1	0	0	1	0	0.546	0.651	76.67378
Kit Kat	1	0	0	0	0	1	0	1	0	0.313	0.511	76.76860
Twix	1	0	1	0	0	1	0	1	0	0.546		
Reese's Miniatures	1	0	0	1	0	0	0	0	0	0.034		
Reese's Peanut Butter cup	1	0	0	1	0	0	0	0	0	0.720		

```
Twix                          0.906   81.64291
Reese's Miniatures            0.279   81.86626
Reese's Peanut Butter cup     0.651   84.18029
```

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

A15.

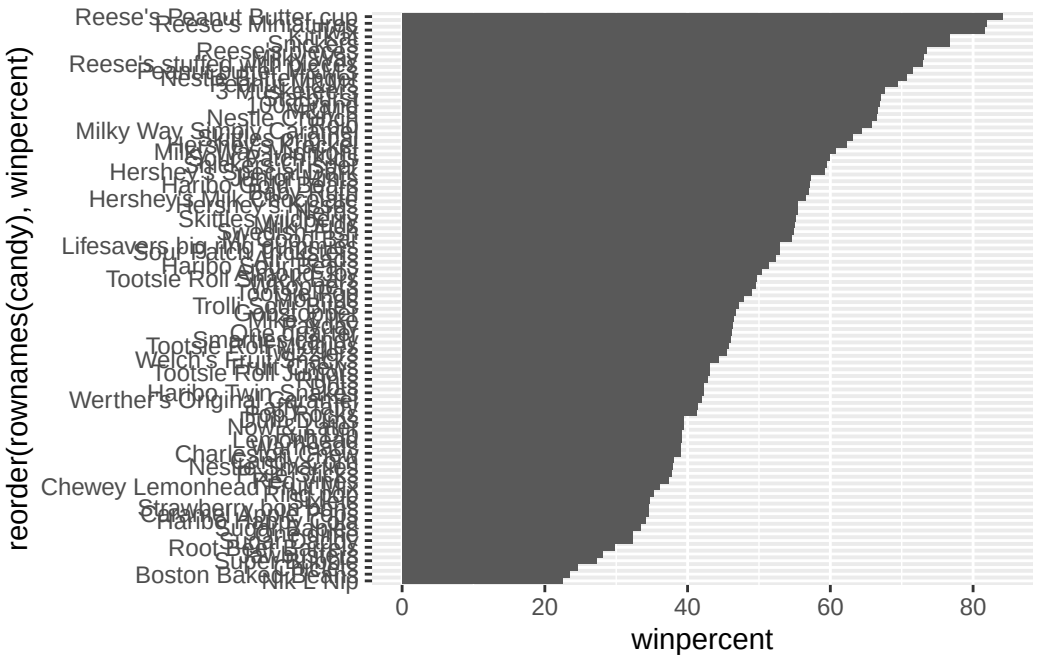
```
library(ggplot2)

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



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

A16.

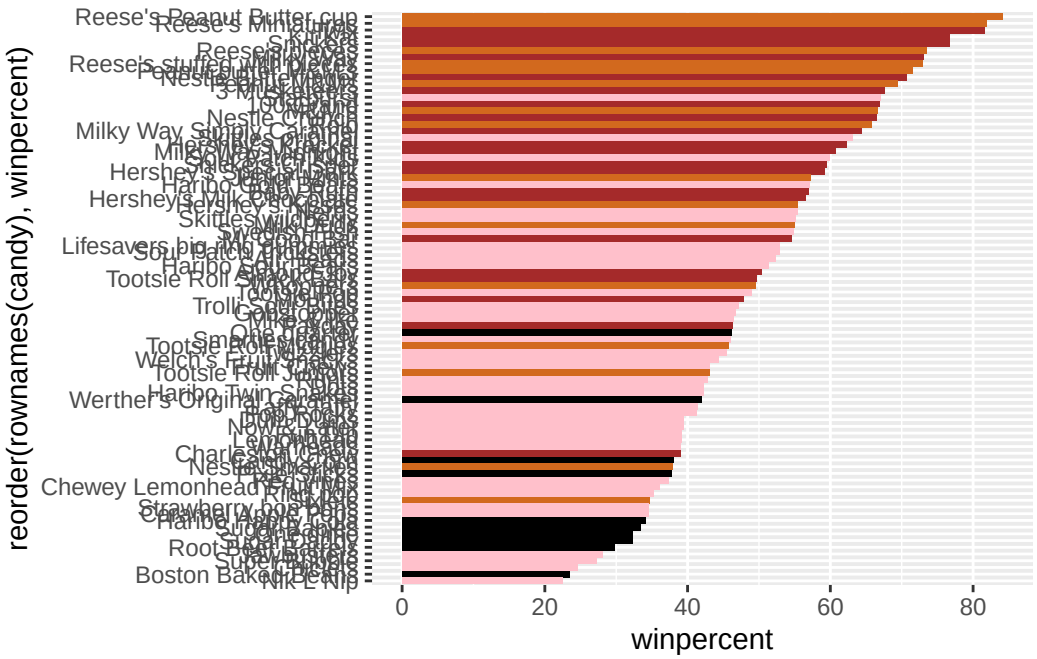
```
library(ggplot2)

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



Let's add color now!

```
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"

# Here are a few color vectors set up for some of our candy columns.
```

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)
```



Q17. What is the worst ranked chocolate candy?

A17. From the graph it is the lowest falling chocolate color bar, which is Sixlets.

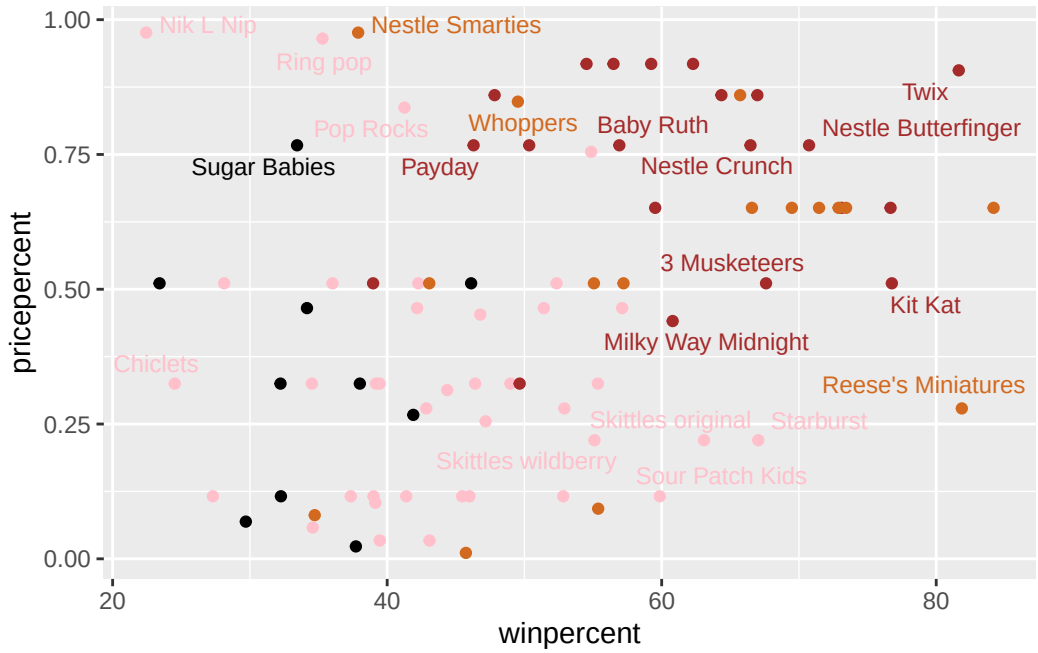
Q18. What is the best ranked fruity candy?

A18. The best ranked fruity candy is Starburst (top pink bar, taken from graph)!

Taking a look at pricepercent

```
# install.packages("ggrepel")
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=3.3, max.overlaps = 5)
```

```
Warning: ggrepel: 65 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?

A19. Reese's Miniatures! They sit around a pricepercent of 0.3 meanwhile being at a winpercent of about 83!

Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

A20. The top 5 most expensive candy types are Nik L Nip tied in pricepercent with Nestle Smarties, Ring pop next, Hershey's Krackel, then Hershey's Milk Chocolate, with Nik L Nip being the least popular with the lowest winpercent (22.44534) amongst the group.

```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

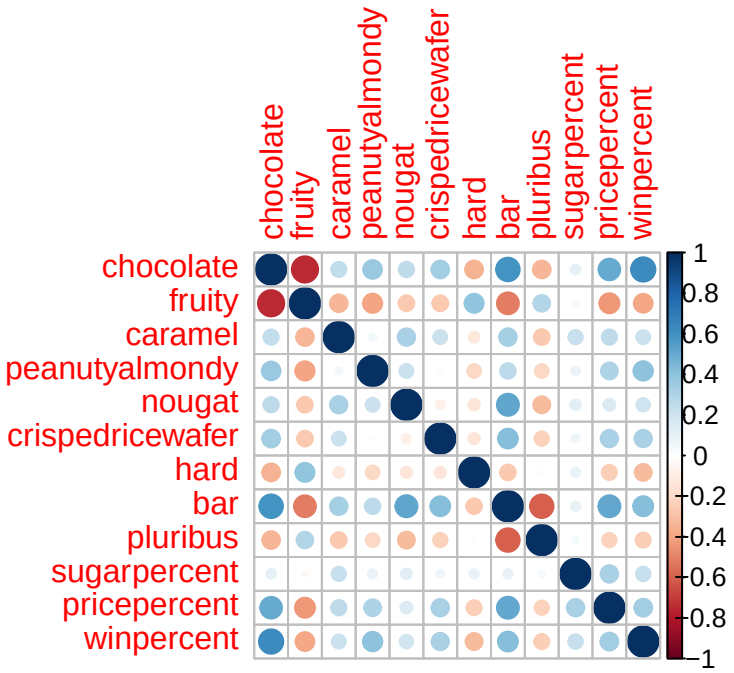
```
                      pricepercent winpercent
Nik L Nip                    0.976   22.44534
Nestle Smarties              0.976   37.88719
Ring pop                     0.965   35.29076
Hershey's Krackel            0.918   62.28448
Hershey's Milk Chocolate     0.918   56.49050
```

Exploring the correlation structure

```
# install.packages("corrplot")
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



Woah!

Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

A22. The fruity and chocolate variables are anti-correlated, having minus values colored red in the plot. Which is a shame because what do you know about the Meiji Choco Gummy's (they might sell at market!)

Q23. Similarly, what two variables are most positively correlated?

A23. The two variables most positively correlated are the variables against itself (e.g. bar x bar, nougat x nougat, and chocolate x chocolate).

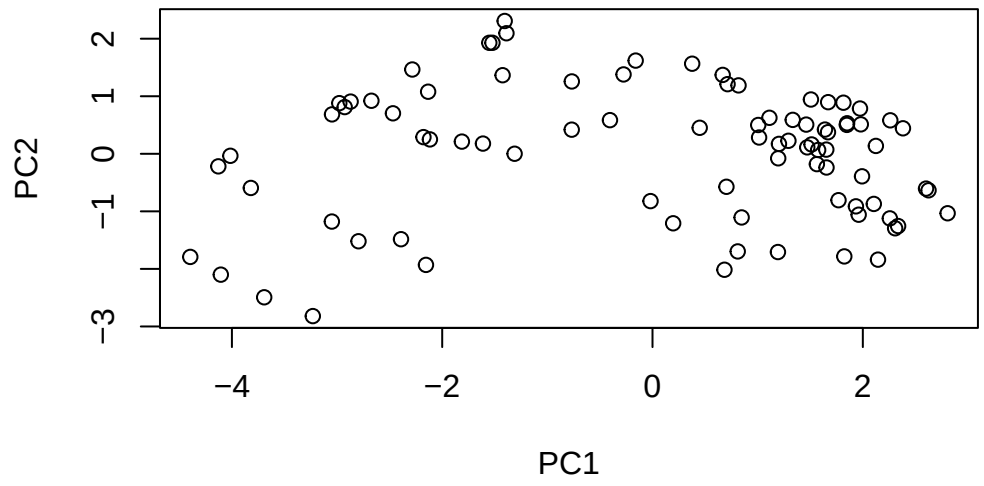
PCA

```
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
```

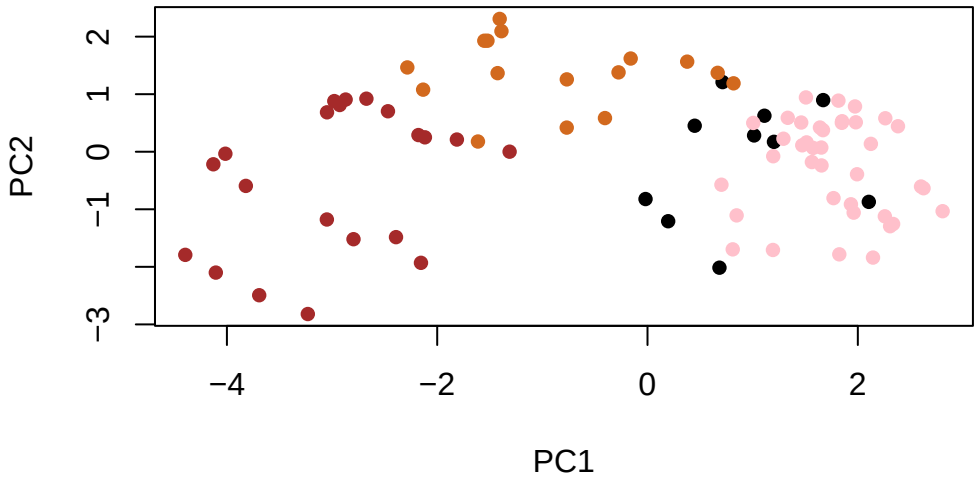
Now, we plot >:)

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



Now, some color!

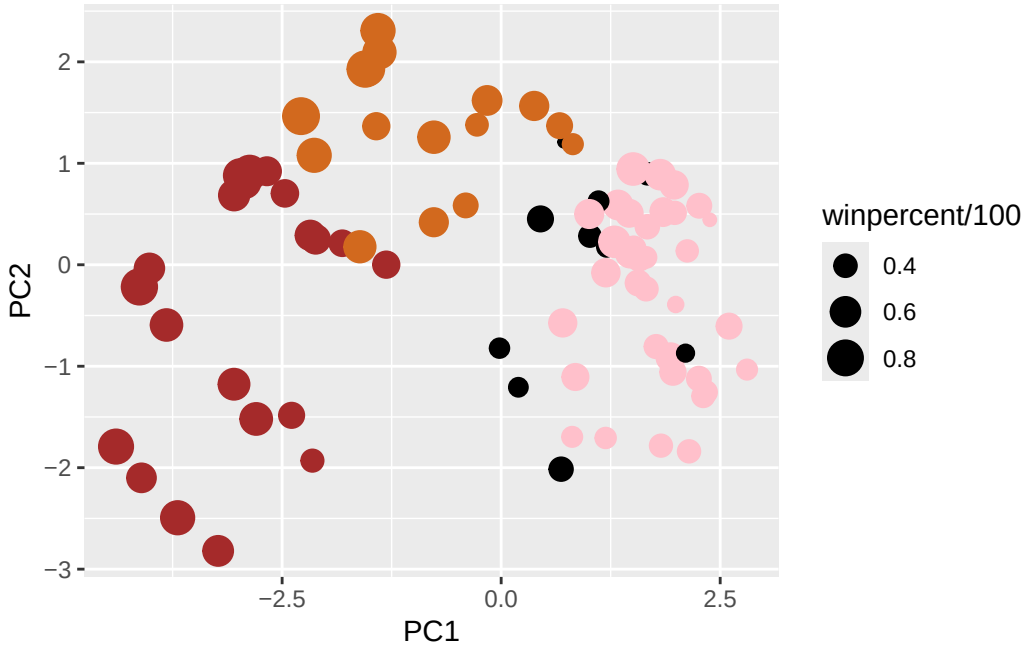
```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



Let's try with ggplot, which prefers df inputs that have sep columns for each aes for the plot.

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

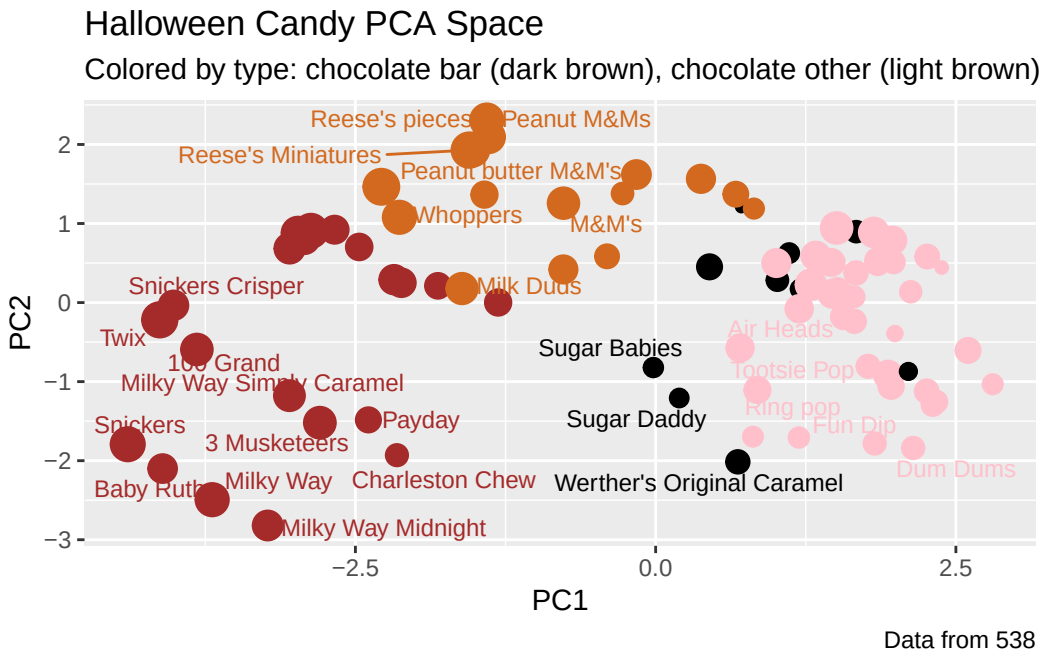
```
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



Now we add the text_repel for candy names and other labels.

```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
  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")
```

```
Warning: ggrepel: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



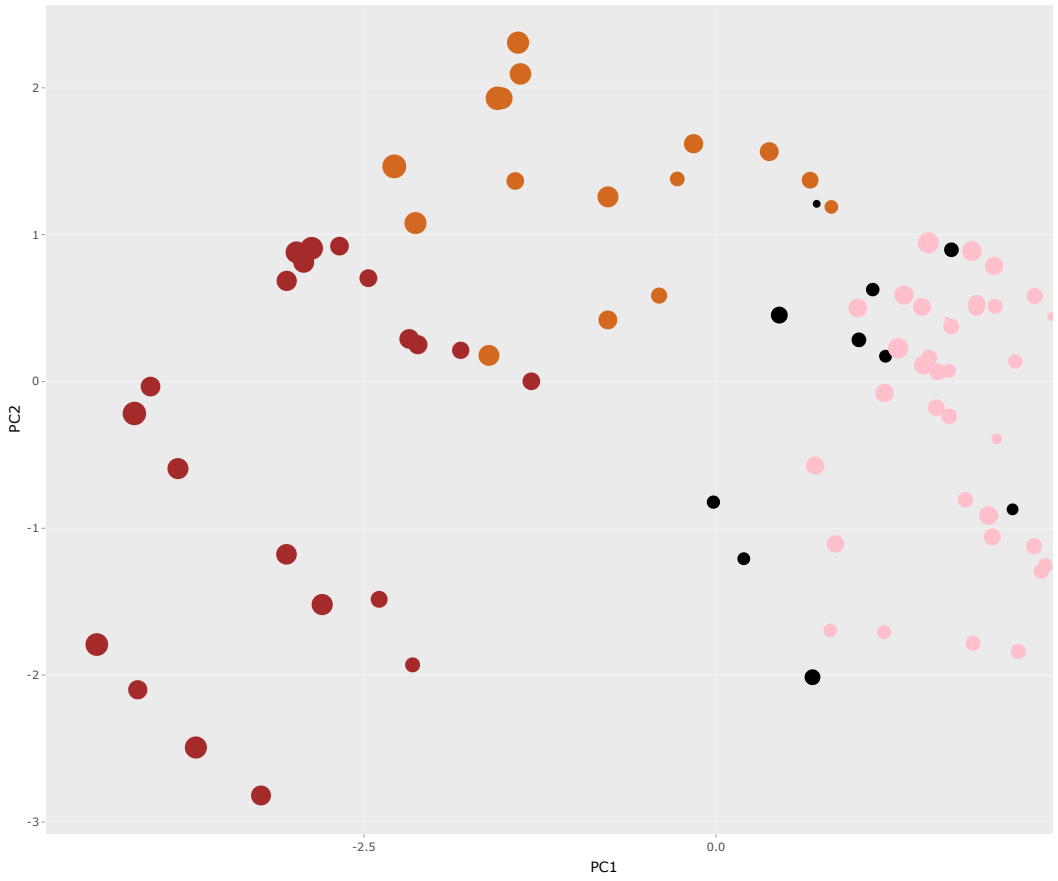
Let's try using this ggplot (p) with plotly to generate an interactive plot.

```
# install.packages("plotly")
library(plotly)
```

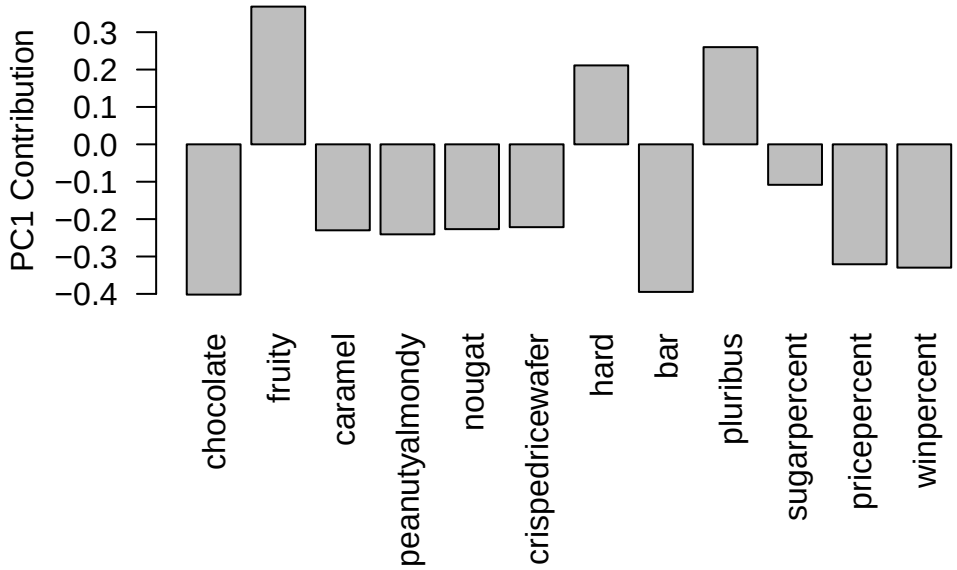
```
Attaching package: 'plotly'

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)
```

file:////private/var/folders/l3/0w1v1h157qj81sz_k6f27glm0000gn/T/RtmpH7sHev/file7b2d252370b6,


2
1
0
-1
-2
-3
PC2
-2.5
0.0
PC1

```r
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?

A24. Fruity, hard, and pluribus are picked up strongly by PC1 in the positive, as shown in the barplot. These make sense to me, as the candy market for multiple little fruity hard candy in one package is huge (Skittles are so delicious). This might suggest that PC1 may represent a contrast between fruity, hard, multi-piece candies and others types!