Class 10: Halloween Mini-Project

Ellice Wang

2025-02-04

Table of contents

Today we will examine data from 538 on common Halloween candy. We will use ggplot, dplyr, and PCA to make sense of this multivariate dataset.

```r
# load in libraries
library(ggplot2)
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
```

1. Importing candy data

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

candy <- read.delim(url, sep = ",")
```

```
candy_file <- "candy-data.csv"

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
```

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

There are 85 different candies in this dataset.

```
nrow(candy)
```

```
[1] 85
```

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

There are 38 fruity candy types in this dataset.

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

```
[1] 38
```

2. What is your favorite candy?

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

My favorite candy is Reese’s Peanut Butter cups. It has a winpercent value of **84.18029**.

```r
candy["Reese's Peanut Butter cup",]$winpercent
```

```
[1] 84.18029
```

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

The win percent value for Kit kat is **76.7686**.

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

```
[1] 76.7686
```

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

The win percent value for Tootsie Roll Snack Bars is **49.6535**.

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

```
[1] 49.6535
```

```r
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?

It seems that winpercent row is different than the majority of the other columns in the dataset. The other columns range from 0 to 1 while winpercent goes from 14%-84%.

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

It represents whether or not the candy contains chocolate (1) or if it does not (0).

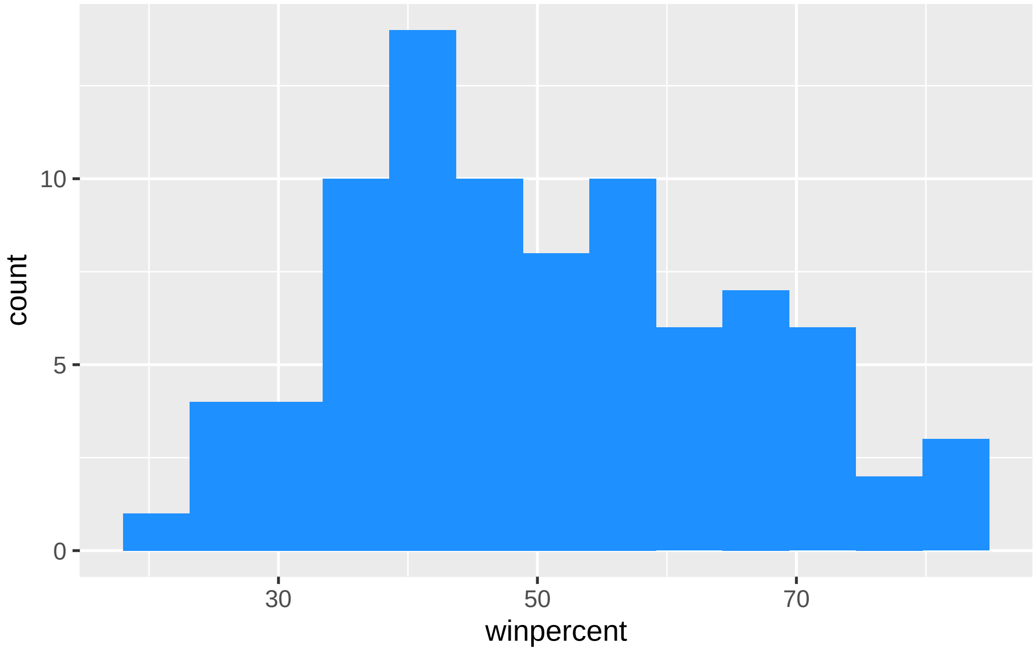
```
head(candy)
```

```
             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
```

Q8. Plot a histogram of winpercent values

```r
ggplot(candy, aes(x=winpercent)) +
  geom_histogram(bins=13, fill="dodgerblue")
```



Q9. Is the distribution of winpercent values symmetrical?

No the distribution is not symmetrical. The data is slightly skewed towards the left side of the graph, or a lower win percent value.

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

The center of the distribution is around 50%. The median is below 50% but the mean is at 50.32%.

```
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.14   47.83   50.32   59.86   84.18 
```

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

On average, the choclate candy is ranked higher than the fruit candy.

```
chocolate_win <- candy$winpercent[(candy$chocolate) == 1]
mean.choc_win <- mean(chocolate_win)
fruit_win <- candy$winpercent[as.logical(candy$fruity)]
mean.fruit_win <- mean(fruit_win)

paste("chocolate:", mean.choc_win, "fruit:", mean.fruit_win, sep=" ")
```

```
[1] "chocolate: 60.9215294054054 fruit: 44.1197414210526"
```

> Q12. Is this difference statistically significant?

This different is statistically significant. The p-value of the chocolate and fruit data is < 0.05 which suggests that the difference is statistically significant.

```
t.test(chocolate_win, fruit_win)
```

```

	Welch Two Sample t-test

data:  chocolate_win and fruit_win
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 
```

3. Overall Candy Rankings

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

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

```
head(arrange(candy, winpercent), 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
```

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

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

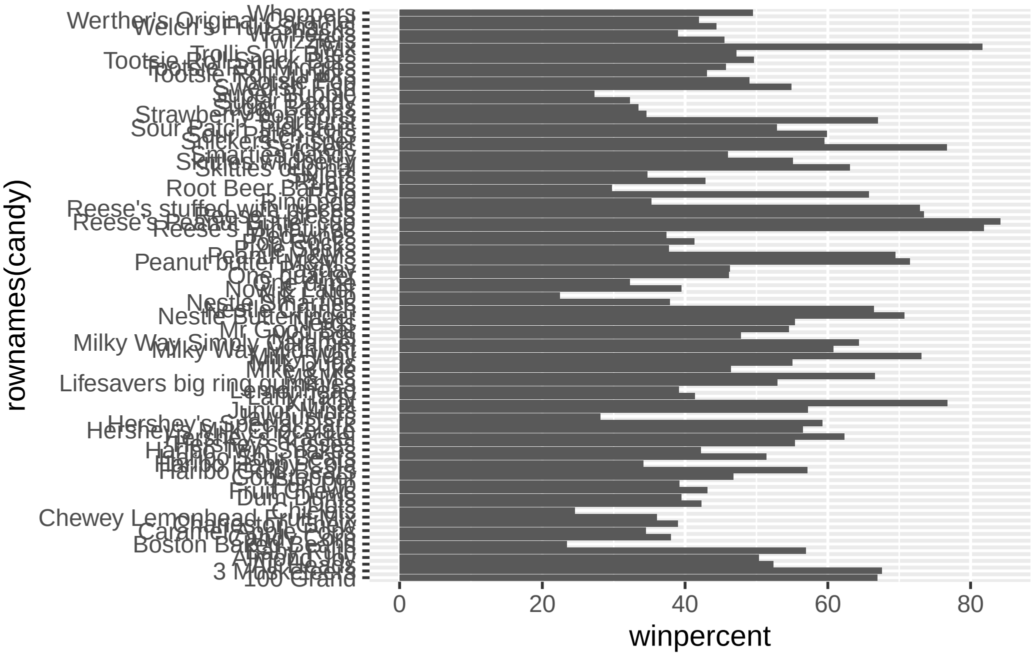
```
head(arrange(candy, desc(winpercent)), 5)
```

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

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

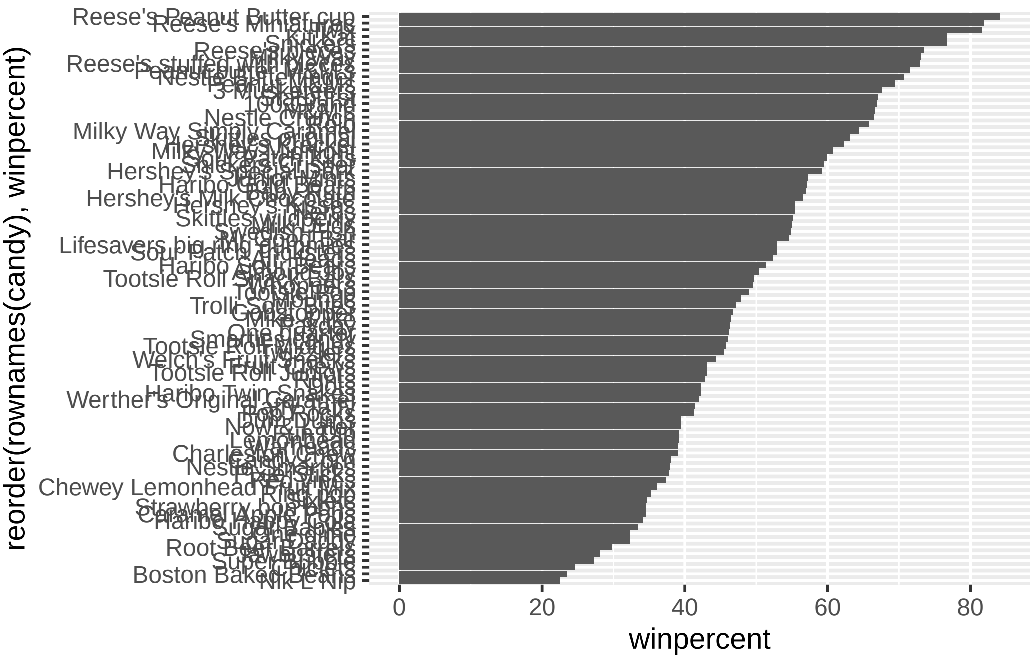
Q15. Make a first barplot of candy ranking based on winpercent values.

```r
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col(position = "dodge")
```



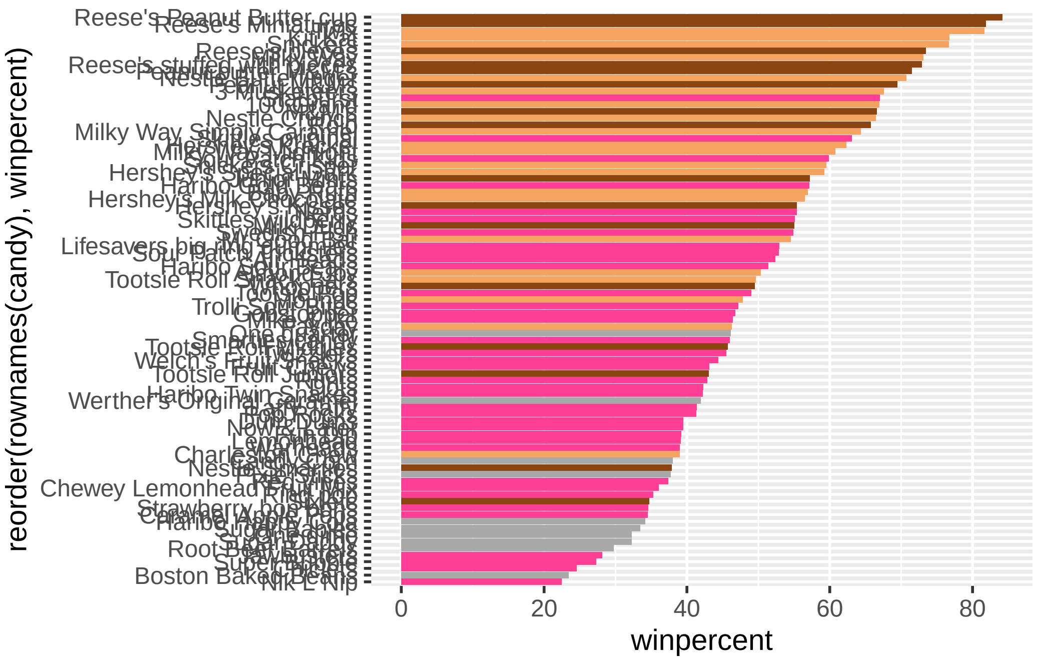
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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



```
my_cols=rep("darkgrey", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate4"
my_cols[as.logical(candy$bar)] = "sandybrown"
my_cols[as.logical(candy$fruity)] = "violetred1"
```

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



Q17. What is the worst ranked chocolate candy?

The worst ranked chocolate candy are Sixlets.

Q18. What is the best ranked fruity candy?

The best ranked fruit candy are Starbursts

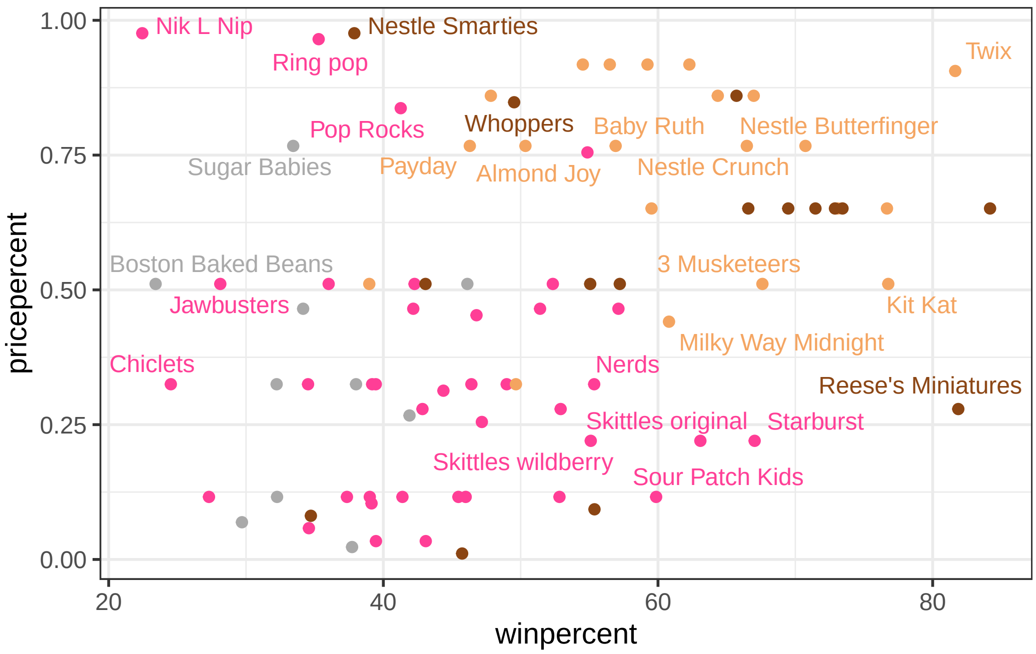
4. Taking a look at pricepercent

```
# package to avoid over-plotting
library(ggrepel)
```

```
Warning: package 'ggrepel' was built under R version 4.4.1
```

```
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 6) +
  theme_bw()
```

```
Warning: ggrepel: 61 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?

Reese's Miniatures, Starburst, Sour Patch Kids, and Skittles original candy are highly ranked candy that do not cost a lot of money.

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

The top. 5 most expensive candy types in the dataset are Nik L Nip, Nestle Smarties, Ring pop, Mr Good Bar, and Hershey's Milk Chocolate. The least popular candy is Nik L Nip.

```r
head(arrange(candy, desc(pricepercent), winpercent))
```

```
                         chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                        0      1       0              0      0
Nestle Smarties                  1      0       0              0      0
Ring pop                         0      1       0              0      0
Mr Good Bar                      1      0       0              1      0
Hershey's Milk Chocolate         1      0       0              0      0
```

```
Hershey's Special Dark           1      0       0              0      0
                         crispedricewafer hard bar pluribus sugarpercent
Nik L Nip                               0    0   0        1        0.197
Nestle Smarties                         0    0   0        1        0.267
Ring pop                                0    1   0        0        0.732
Mr Good Bar                             0    0   1        0        0.313
Hershey's Milk Chocolate                0    0   1        0        0.430
Hershey's Special Dark                  0    0   1        0        0.430
                         pricepercent winpercent
Nik L Nip                       0.976   22.44534
Nestle Smarties                 0.976   37.88719
Ring pop                        0.965   35.29076
Mr Good Bar                     0.918   54.52645
Hershey's Milk Chocolate        0.918   56.49050
Hershey's Special Dark          0.918   59.23612
```

5. Exploring the correlation structure

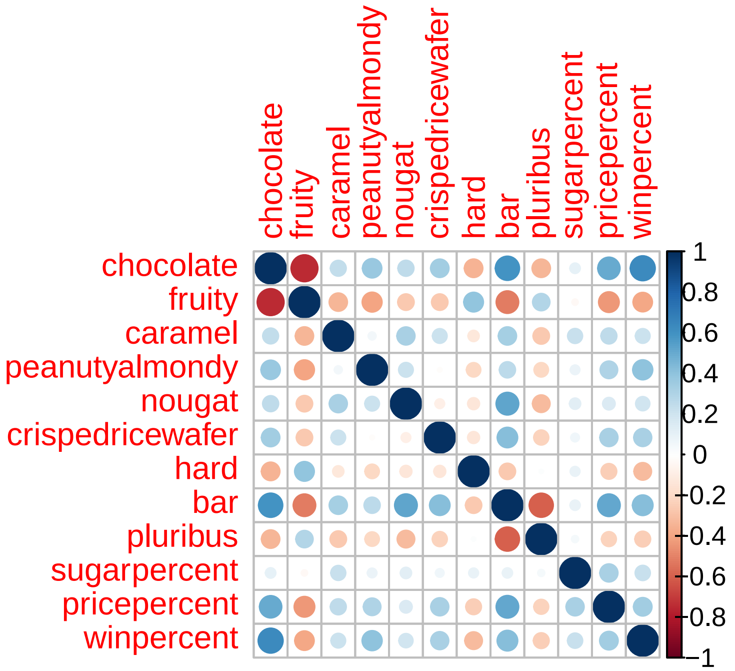
Now that we have explored the dataset a little, we will see how the variables interact with one another.

```
library(corrplot)
```

```
Warning: package 'corrplot' was built under R version 4.4.1
```

```
corrplot 0.95 loaded
```

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



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

The fruity and chocolate variables are the most anti-correlated, which means that not a lot of fruity chocolate candies exist.

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

Chocolate and winpercent, and chocolate and bar are very positively correlated.

6. Principal Component Analysis

Let's apply PCA using the 'prcom()' function to our candy dataset remembering to set the **scale=TRUE** argument.

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

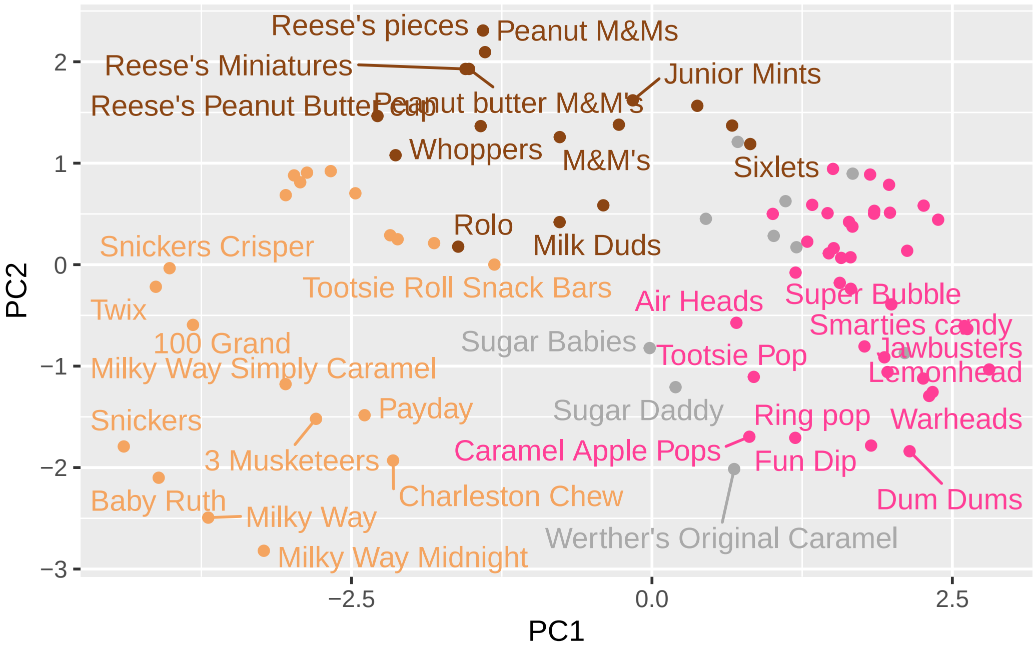
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
```

```
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
```

Let's plot our main results as our PCA "score plot"

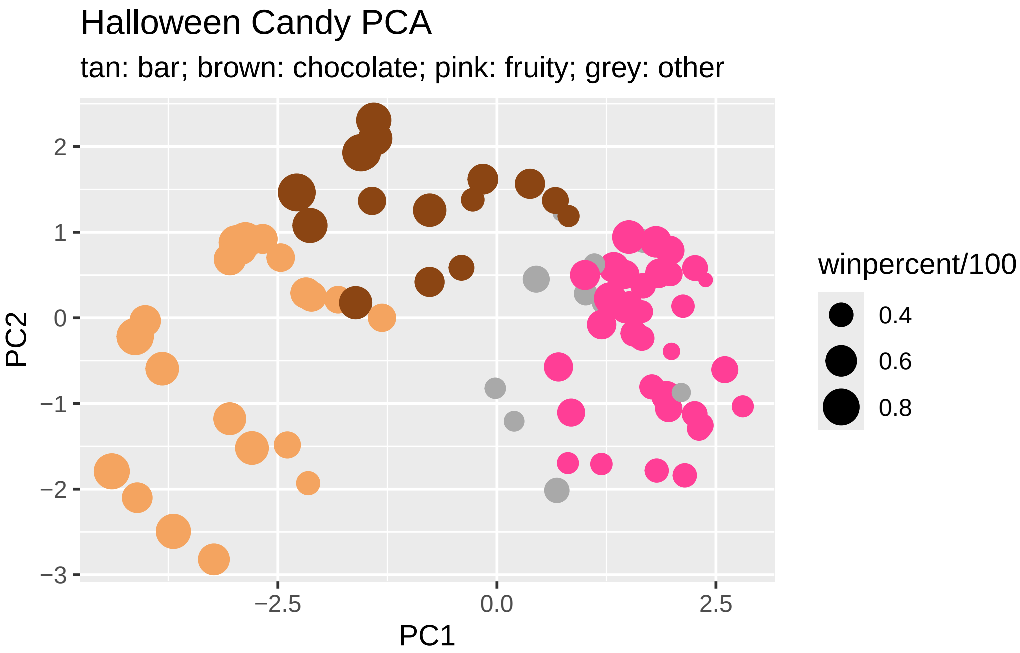
```
p <- ggplot(pca$x) +
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols)
p
```

```
Warning: ggrepel: 48 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



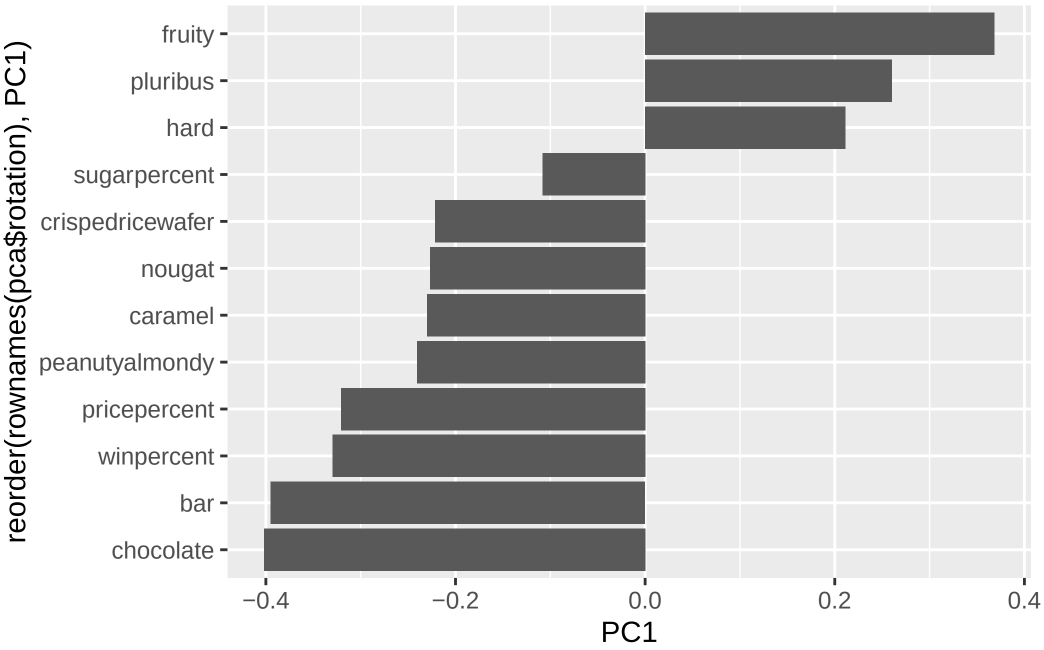
Combine PCA data and candy data

```
pcandy <- cbind(candy, pca$x[,1:3])
p <- ggplot(pcandy) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(pcandy),
            label=rownames(pcandy)) +
        geom_point(col=my_cols) +
  labs(title="Halloween Candy PCA", subtitle = "tan: bar; brown: chocolate; pink: fruity; gre
p
```



```
# library(plotly)
# ggplotly(p)
```

```
ggplot(pca$rotation) +
  aes(PC1, reorder(rownames(pca$rotation), PC1)) +
  geom_col()
```



> Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

Fruity, pluribus, and hard are picked up strongly by PC1 in the positive direction. This makes sense to me since a lot of the fruity candy tend to be in a bag with multiple of them and are generally hard candies. The candy that comes to mind are nerds.