Class08 halloween

Neva Olliffe PID A69026930

1. Import the data

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

candy = read.csv(url(candy_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?

```
dim(candy)
```

```
[1] 85 12
```

85 types of candy.

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

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

```
[1] 38
```

38 types of candy are fruity.

#2. What's your favorite candy?

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

```
[1] 81.64291
```

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

```
[1] 76.7686
```

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

```
[1] 49.6535
```

Twix wins 81.6% of the time.

```
#library("skimr")
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?

Win percent looks like a very different scale.

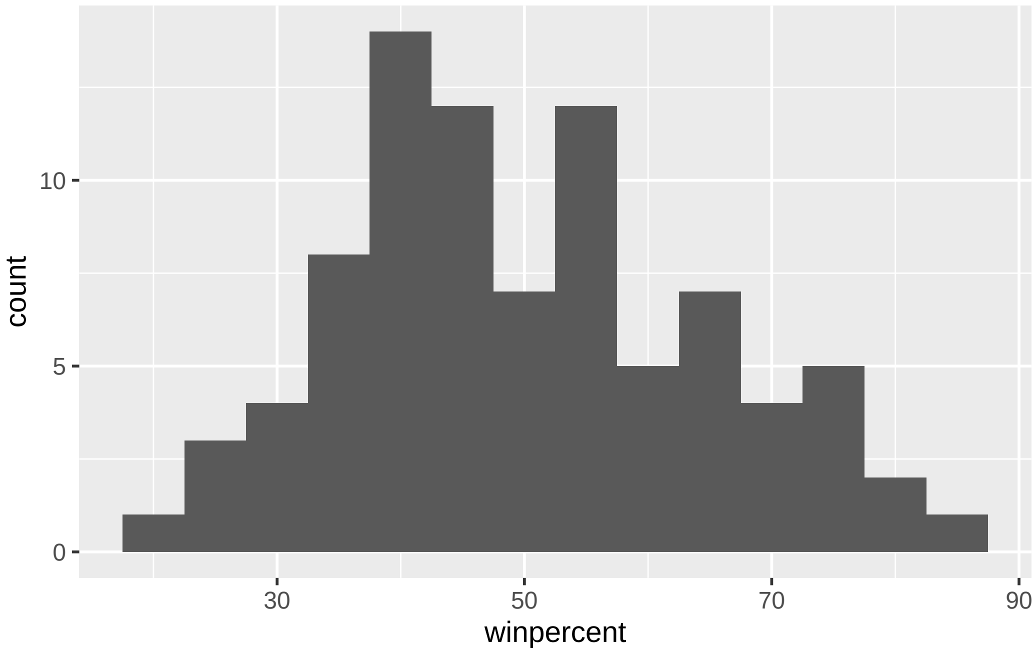
Q7. What do you think a zero and one represent for the candy$chocolate column?

A zero stands in for `False` and a 1 for `True`.

Q8. Plot a histogram of winpercent values

```r
library(ggplot2)

ggplot(candy, aes(x = winpercent)) +
  geom_histogram(binwidth = 5)
```



Q9. Is the distribution of winpercent values symmetrical?

It is slightly skewed right.

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

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

```
[1] 47.82975
```

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

```
[1] 50.31676
```

The center of the distribution is just slightly below 50% if you use the median, or just above if you use the mean.

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

```r
choc_candy <- candy[as.logical(candy$chocolate), "winpercent"]
fruit_candy <- candy[as.logical(candy$fruity), "winpercent"]

mean(choc_candy)
```

```
[1] 60.92153
```

```r
mean(fruit_candy)
```

```
[1] 44.11974
```

Chocolate candy is ranked higher on average than fruit candy.

Q12. Is this difference statistically significant?

```r
t.test(choc_candy, fruit_candy)
```

```

	Welch Two Sample t-test

data:  choc_candy and fruit_candy
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

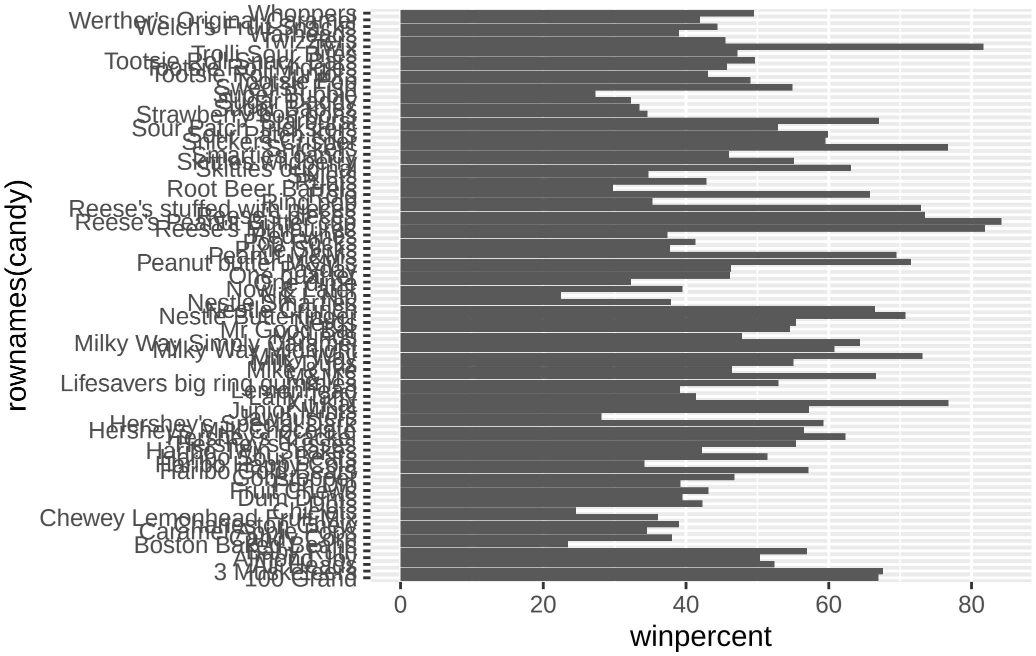
Q13. What are the five least liked candy types in this set?

```
head(candy[order(candy$winpercent),], n=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
```

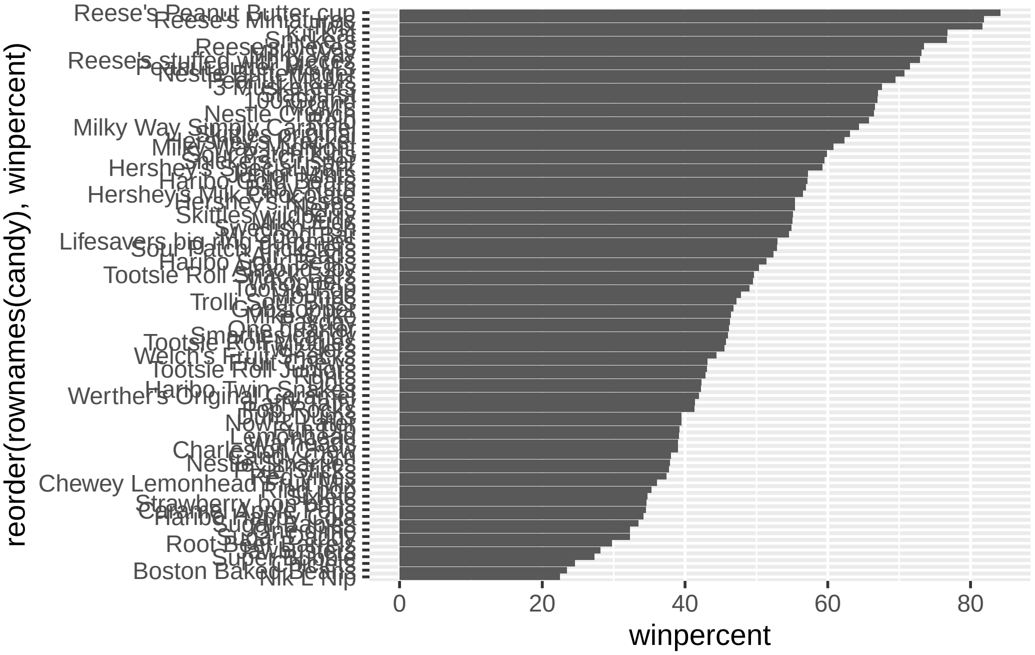
Q15. Make a first barplot of candy ranking based on winpercent values.

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



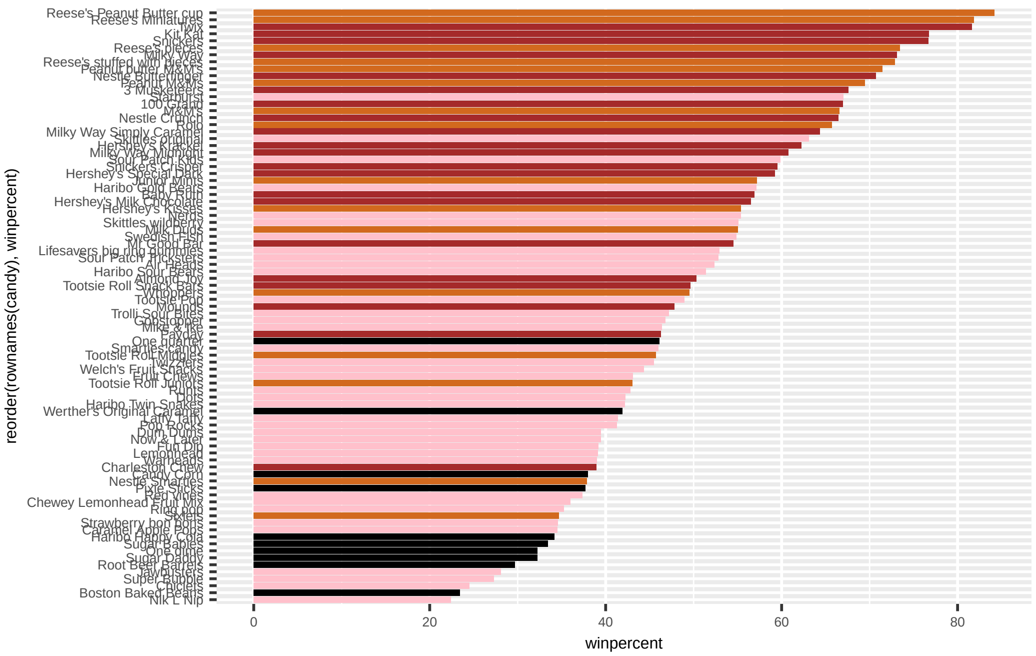
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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



```
my_cols=rep("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_cols) +
  theme(text = element_text(size=6))
```



Q17. What is the worst ranked chocolate candy?

Sixlets.

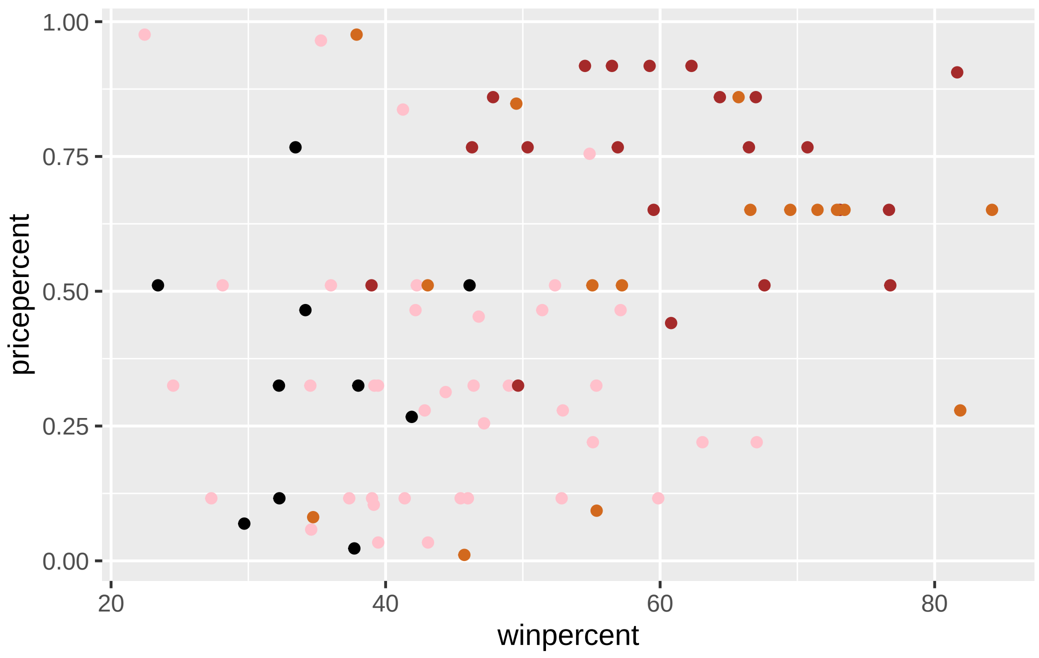
Q18. What is the best ranked fruity candy?

Starburst.

Taking a look at pricepoint.

```
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)
```

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?

```
row.names(candy)[which.max(candy$winpercent/candy$pricepercent)]
```

```
[1] "Tootsie Roll Midgies"
```

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

```
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
```

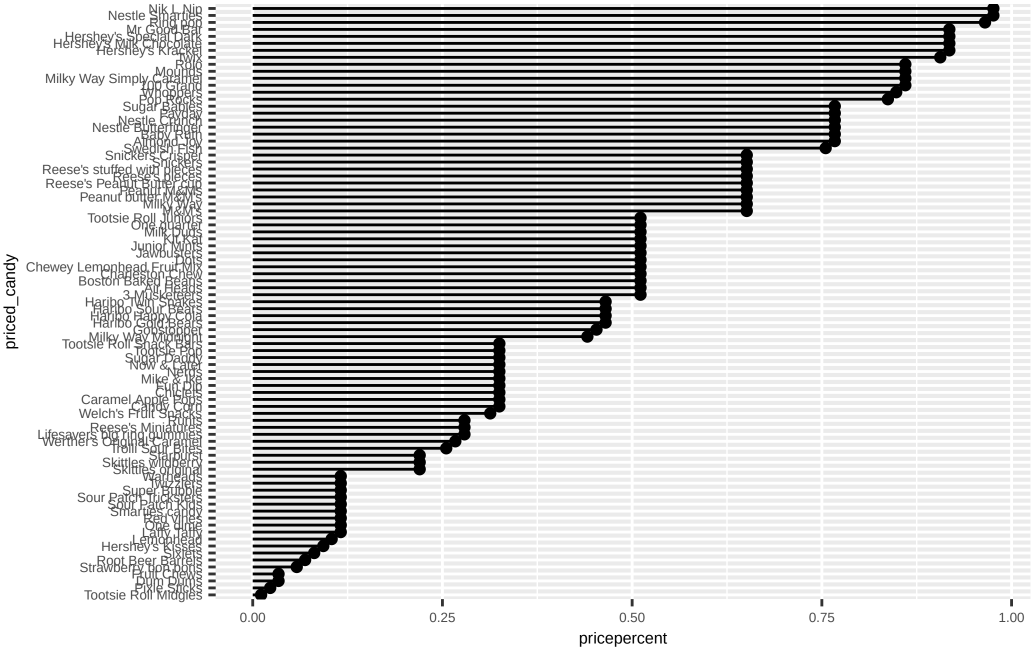
```
Ring pop                         0.965   35.29076
Hershey's Krackel                0.918   62.28448
Hershey's Milk Chocolate         0.918   56.49050
```

Nik L Nip is the most expensive candy and the least popular of the top 5 most expensive candies.

Q21. Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping geom_col() for geom_point() + geom_segment().

```r
priced_candy <- reorder(rownames(candy), candy$pricepercent)

ggplot(candy, aes(x = pricepercent, priced_candy)) +
  geom_point() +
  geom_segment(aes(x = pricepercent, xend = 0, yend = priced_candy,
                   y = priced_candy)) +
    theme(text = element_text(size=6))
```

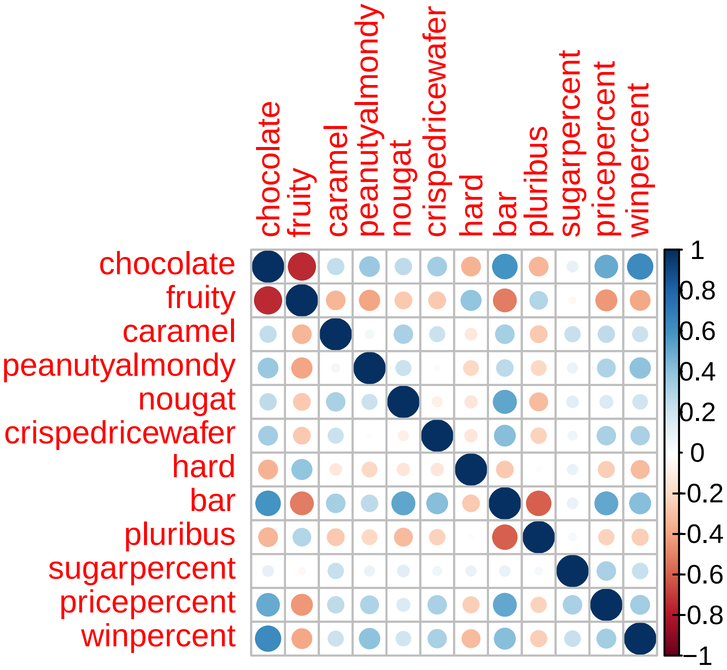


5 Exploring the correlation structure

```
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)?

Fruity & chocolate.

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

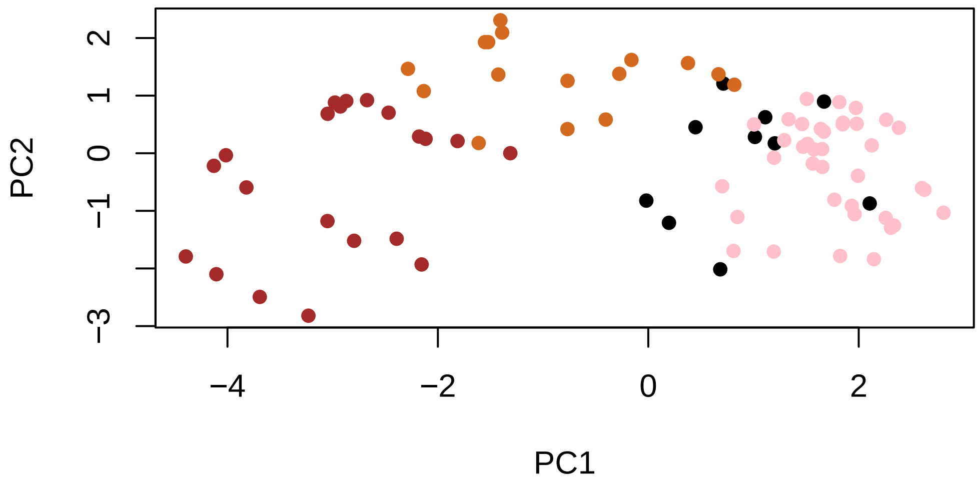
chocolate & winpercent.

6. Principal Component Analysis

```
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
```

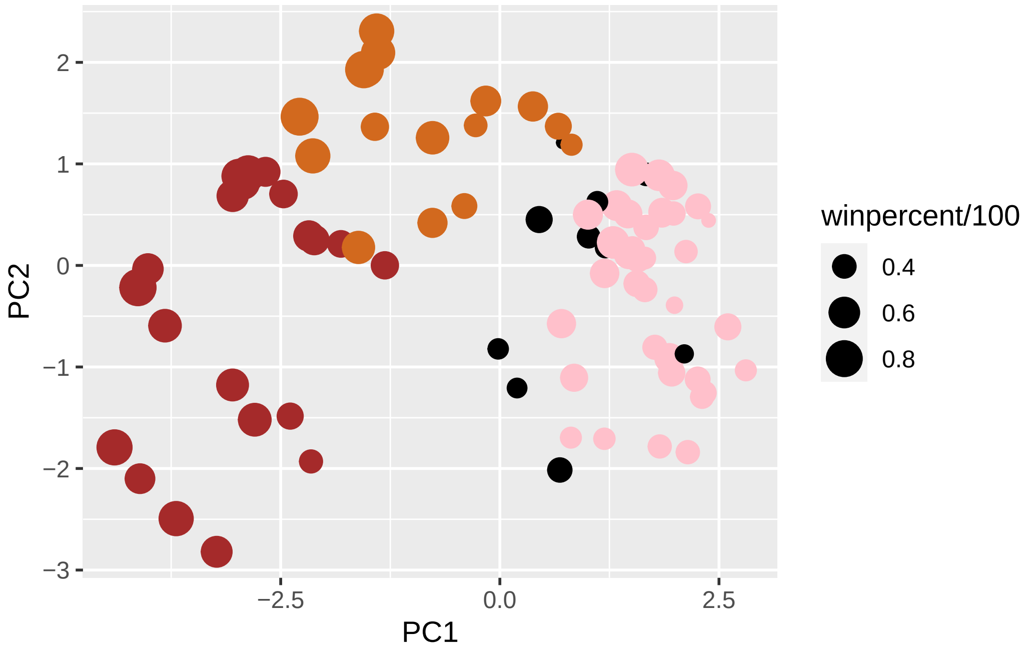
```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



```
# 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
```

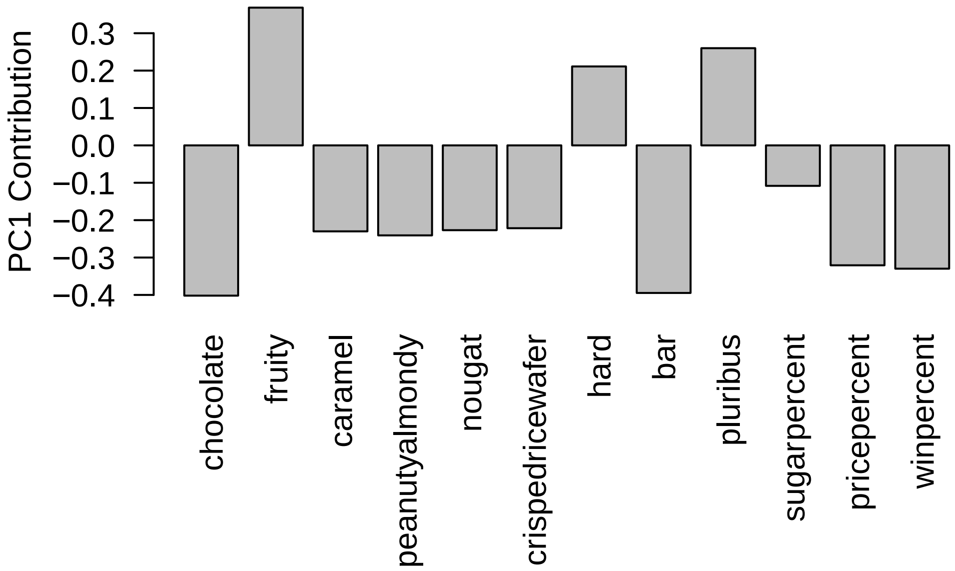


```
library(ggrepel)

#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 brow
       #caption="Data from 538")
```

```
#library(plotly)
#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?

Fruity, hard, and pluribus are picked up strongly. This makes sense because many fruity candies come in multiple packs (skittles, starbursts, etc) which is less common for chocolate candies.