Class09: Halloween

Nathaniel Lightle (A16669288)

Today is Halloween and we will apply lots of the analysis methods and R graphics appraoches to find out all about typical Halloween candy.

Importing candy data

```
candyphile = read.csv("candy-data.csv", row.names = 1)
head(candyphile)
```

```
             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 data set?

```
nrow(candyphile)
```

```
[1] 85
```

There are 85 different types of candy in the data set

[Q2] How many fruity candy types are in the dataset?

```
sum(candyphile[,2])
```

```
[1] 38
```

There are 38 fruity candy types in the dataset

What is your favorite candy?

[Q3] What is your favorite candy in the dataset and what is it's win percent?

My favorite candy is Werther's Original's Caramel

```
candyphile["Werther's Original Caramel", "winpercent"]
```

```
[1] 41.90431
```

The win percent for Werther's Original Caramel is 41.90431%

[Q4] What is the winpercent value for "Kit Kat"?

```
candyphile["Kit Kat", "winpercent"]
```

```
[1] 76.7686
```

The win percent for Kit Kat is 76.7686%

[Q5] What is the winpercent value for "Tootsie Roll Snack Bars"?

```
candyphile["Tootsie Roll Snack Bars", "winpercent"]
```

```
[1] 49.6535
```

The win percent for Tootsie Roll Snack Bars is 49.6535%

Trying the `skim()` function

```
#install.packages("skimr")
library("skimr")
skim(candyphile)
```

Table 1: Data summary

Name	candyphile
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?

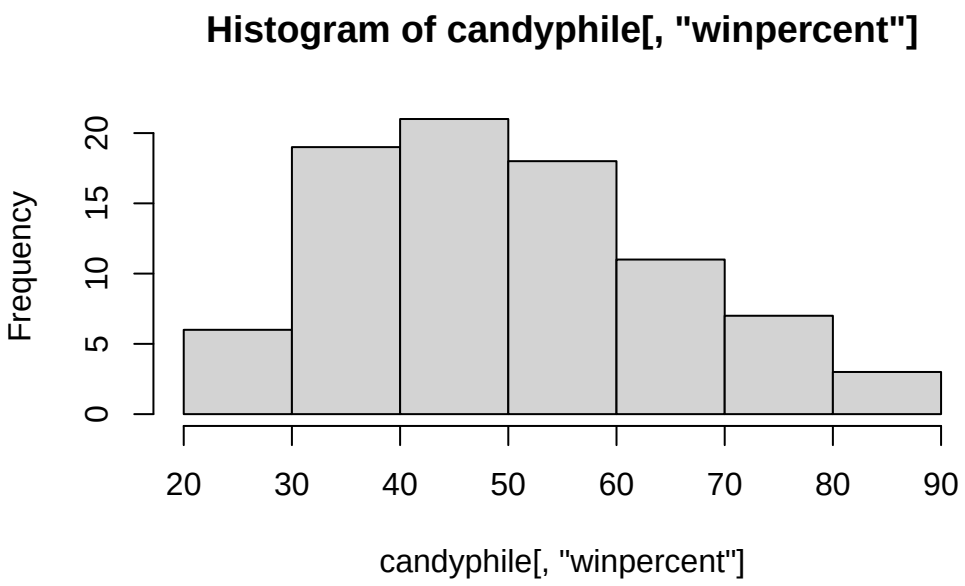
The win percent is on a different scale because it ranges from 0-100%

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

A 0 means false and a 1 means true

[Q8] Plot a histogram of winpercent values

```
hist(candyphile[,"winpercent"])
```



[Q9] Is the distribution of winpercent values symmetrical?

No the distribution is skewed to the left

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

The center is below 50%

[Q11] On average is chocolate candy higher or lower ranked than fruity candy?

```
mean(candyphile$winpercent[as.logical(candyphile$chocolate) == T])
```

```
[1] 60.92153
```

```
mean(candyphile$winpercent[as.logical(candyphile$fruity) == T])
```

```
[1] 44.11974
```

On average chocolate candy is higher ranked than fruity candy.

[Q12] Is this difference statistically significant?

```
t.test(candyphile$winpercent[as.logical(candyphile$chocolate) == T], candyphile$winpercent
```

```
    Welch Two Sample t-test

data:  candyphile$winpercent[as.logical(candyphile$chocolate) == T] and candyphile$winpercent
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
```

The difference is statistically significant because the p value is teeny tiny

Overall Candy Rankings

Ordering data by winpercent

```
head(candyphile[order(candyphile$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
```

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

The 5 least liked candy types in the dataset are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters

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

```
tail(candyphile[order(candyphile$winpercent),], n=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
Reese's Miniatures              1      0       0              1      0
Reese's 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
Reese's Miniatures                     0    0   0        0        0.034
Reese's 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
Reese's Miniatures             0.279   81.86626
Reese's Peanut Butter cup      0.651   84.18029
```

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

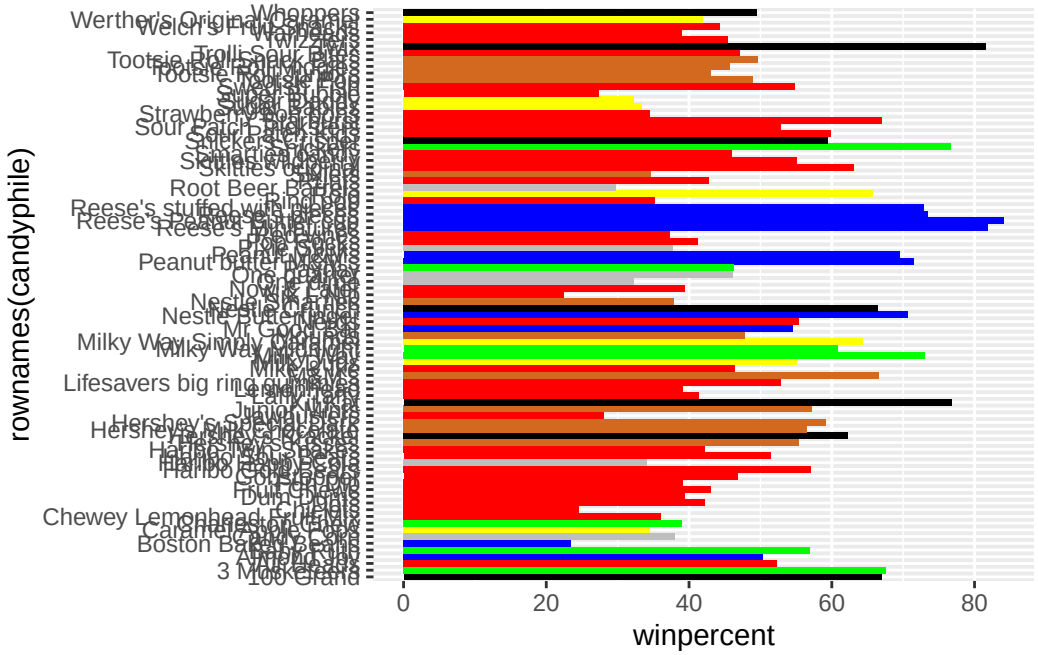
[Q15] Make a first barplot of candy ranking based on winpercent values.

```
mycols <- rep("gray", nrow(candyphile))
#mycols[2:5] <- "red"
mycols[as.logical(candyphile$fruity) == T] <- "red"
mycols[as.logical(candyphile$chocolate) == T] <-"chocolate"
mycols[as.logical(candyphile$caramel) == T] <- "yellow"
```

```
mycols[as.logical(candyphile$peanutyalmondy) == T] <- "blue"
mycols[as.logical(candyphile$nougat) == T] <- "green"
mycols[as.logical(candyphile$crispedricewafer) == T] <- "black"
mycols
```

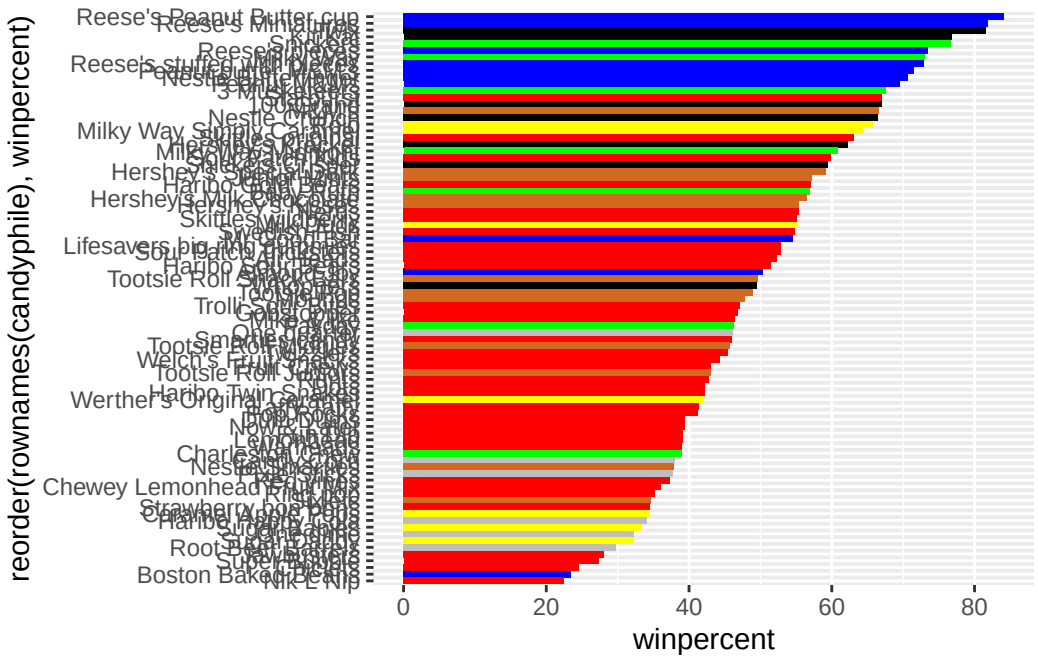
```
 [1] "black"     "green"     "gray"      "gray"      "red"       "blue"
 [7] "green"     "blue"      "gray"      "yellow"    "green"     "red"
[13] "red"       "red"       "red"       "red"       "red"       "red"
[19] "red"       "gray"      "red"       "red"       "chocolate" "black"
[25] "chocolate" "chocolate" "red"       "chocolate" "black"     "red"
[31] "red"       "red"       "blue"      "chocolate" "red"       "yellow"
[37] "green"     "green"     "yellow"    "chocolate" "blue"      "red"
[43] "blue"      "black"     "red"       "red"       "green"     "blue"
[49] "gray"      "red"       "red"       "blue"      "blue"      "blue"
[55] "blue"      "red"       "yellow"    "gray"      "red"       "chocolate"
[61] "red"       "red"       "chocolate" "red"       "green"     "black"
[67] "red"       "red"       "red"       "red"       "yellow"    "yellow"
[73] "red"       "red"       "chocolate" "chocolate" "chocolate" "chocolate"
[79] "red"       "black"     "red"       "red"       "red"       "yellow"
[85] "black"
```

```
library(ggplot2)
ggplot(candyphile) +
  aes(winpercent, rownames(candyphile)) +
  geom_col(fill=mycols)
```



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

```
library(ggplot2)
ggplot(candyphile) +
  aes(winpercent, reorder(rownames(candyphile),winpercent)) +
  geom_col(fill=mycols)
```



[Q17] What is the worst ranked chocolate candy?

Sixlets is the worst ranked chocolate candy

[Q18] What is the best ranked fruity candy?

Starbusts is the best ranked fruity candy

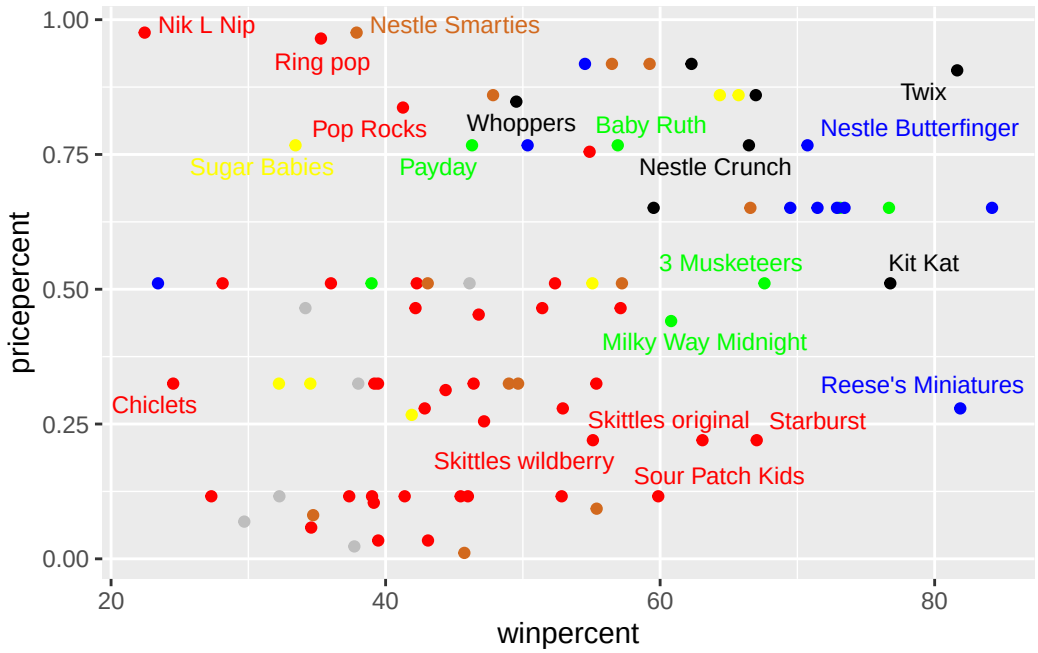
Taking a look at pricepercent

Looking at value Plotting pricepercent vs winpercent

```
library(ggrepel)

ggplot(candyphile) +
  aes(winpercent, pricepercent, label=rownames(candyphile)) +
  geom_point(col=mycols) +
  geom_text_repel(col=mycols, 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?

Reese's Miniatures give you the most bang for your buck

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

```
tail(candyphile[order(candyphile$pricepercent),], n=5)
```

```
                     chocolate fruity caramel peanutyalmondy nougat
Hershey's Special Dark         1      0       0              0      0
Mr Good Bar                    1      0       0              1      0
Ring pop                       0      1       0              0      0
Nik L Nip                      0      1       0              0      0
Nestle Smarties                1      0       0              0      0
                     crispedricewafer hard bar pluribus sugarpercent
Hershey's Special Dark                0    0   1        0        0.430
Mr Good Bar                           0    0   1        0        0.313
Ring pop                              0    1   0        0        0.732
Nik L Nip                             0    0   0        1        0.197
Nestle Smarties                       0    0   0        1        0.267
                     pricepercent winpercent
```

```
Hershey's Special Dark        0.918   59.23612
Mr Good Bar                   0.918   54.52645
Ring pop                      0.965   35.29076
Nik L Nip                     0.976   22.44534
Nestle Smarties               0.976   37.88719
```

The top 5 most expensive candy types in the dataset are Hershey's Special Dark, Mr Good Bar, Ring pop, Nik L nip, and Nestle Smarties.

Exploring the correlation structure

Installing corrplot

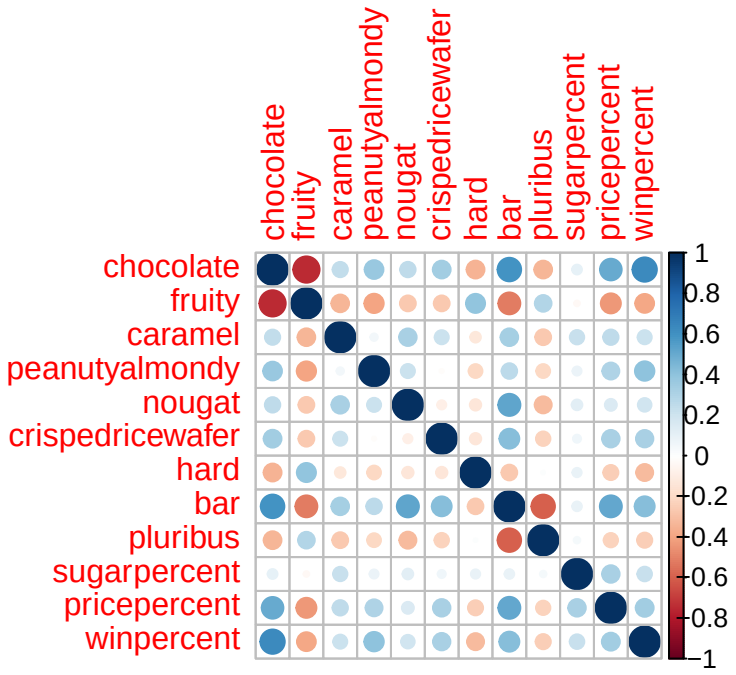
```
#install.packages("corrplot")
```

Using corrplot

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



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

Fruity and chocolate are negatively correlated

[Q23] Similarly, what two variables are most positively correlated?

Winpercent and chocolate are the most highly correlated

PCA Analysis

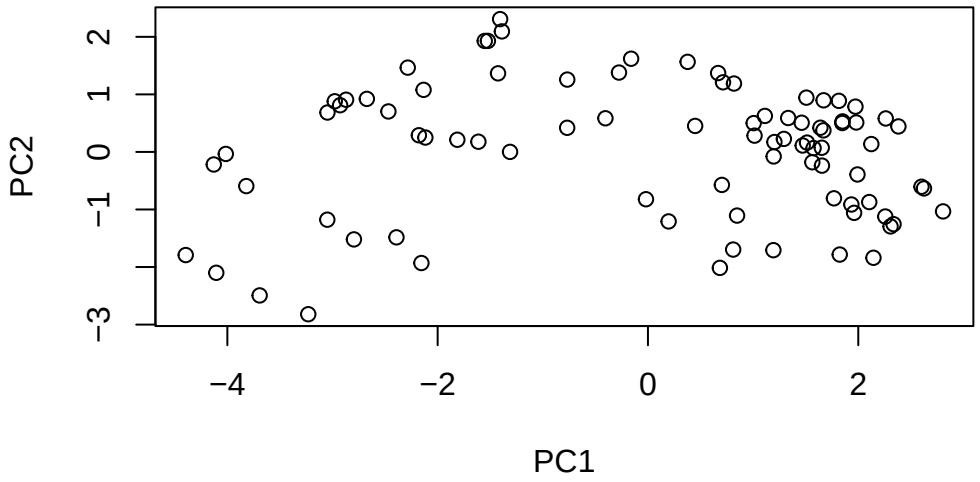
```
pca <- prcomp(candyphile, scale = TRUE)
summary(pca)
```

```
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
```

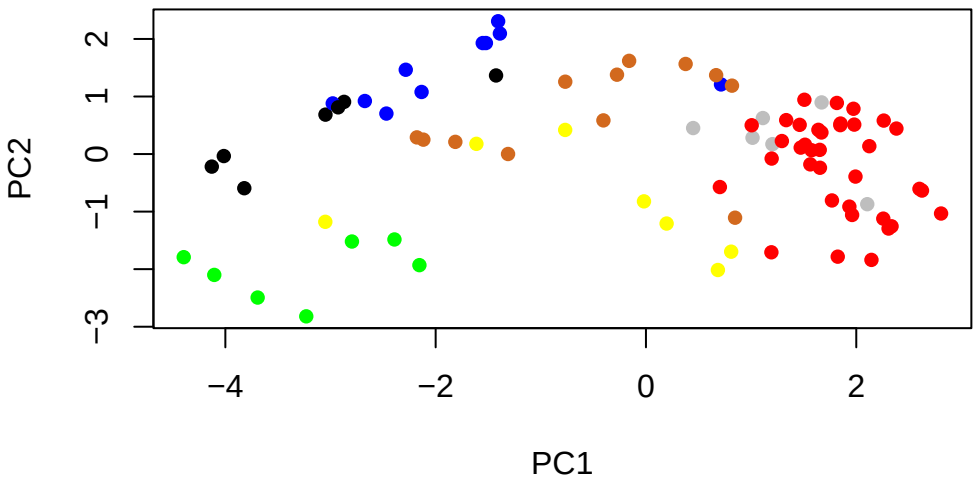
Plotting it

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



Giving it some color

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



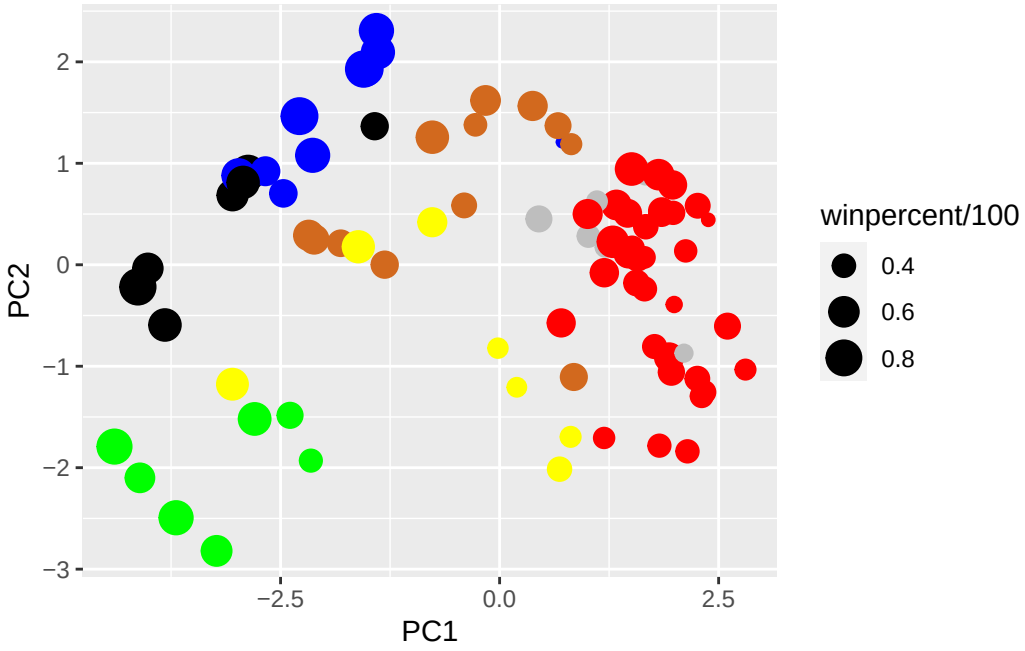
Making new data set of PCA

```
mydata <- cbind(candyphile, pca$x[,1:3])
```

ggPlotting it

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

p
```



Making labels

```
library(ggrepel)

#p + geom_text_repel(size=3.3, col=mycols, 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")
```

```
List of 4
 $ legend.position: chr "none"
 $ title          : chr "Halloween Candy PCA Space"
 $ subtitle       : chr "Colored by type: chocolate bar (dark brown), chocolate other (light
 $ caption        : chr "Data from 538"
 - attr(*, "class")= chr [1:2] "theme" "gg"
 - attr(*, "complete")= logi FALSE
 - attr(*, "validate")= logi TRUE
```

Using plotly

```
#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
```

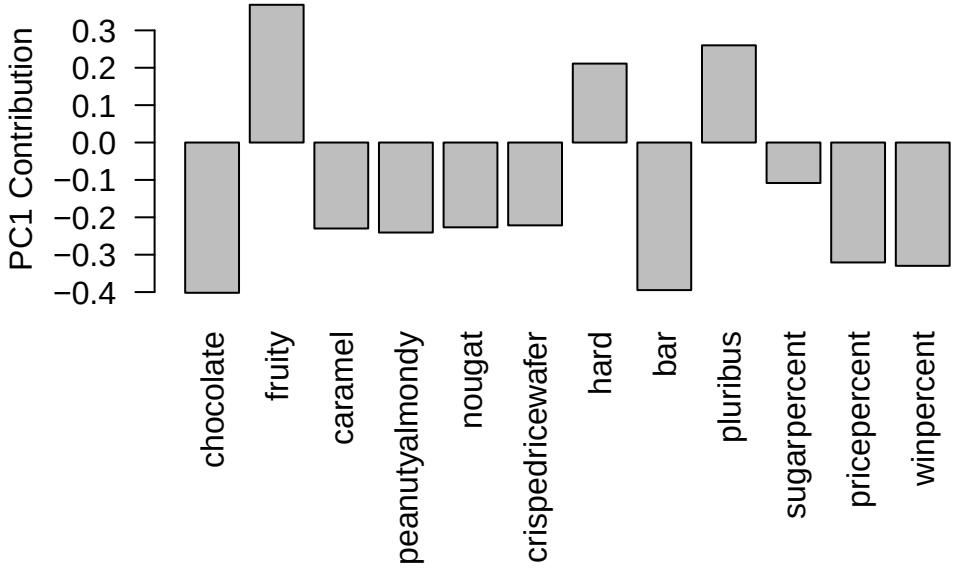
```
The following object is masked from 'package:graphics':

    layout
```

```
#ggplotly(p)
```

Correlation check

```
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 and pluribus are the most strongly correlated in the positive direction. These make sense to me.