class09

Amit Subramanian

install.packages("webshot") webshot::install_phantomjs()

1. Importing candy data

Let's Pull up our CSV file for the candy data set to answer some basic questions on it.

```
candy_file <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-r
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?

```
total_candy_types <- nrow(candy)
total_candy_types
```

```
[1] 85
```

There are 85 different candy types in this dataset.

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

```
fruit_candy_types <- sum(candy$fruity)
fruit_candy_types
```

```
[1] 38
```

There are 38 fruity candy types in the dataset.

2. What is your favorite candy?

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

```
candy["M&M's", ]$winpercent
```

```
[1] 66.57458
```

My favorite candy out of all these options is M&M's, and they have a winpercent value of about 66.57.

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

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

```
[1] 76.7686
```

The winpercent value for Kit Kat is about 76.77.

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

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

```
[1] 49.6535
```

The winpercent value for Tootsie Roll Snack Bars is about 49.65.

Lets try implementing the skimr package to get 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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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?

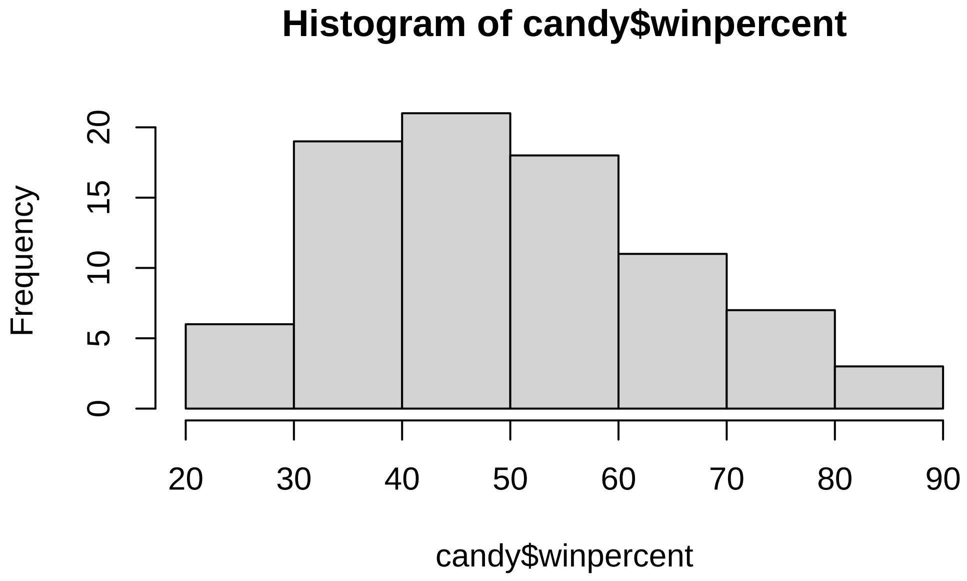
Most variables are under the basis of "yes/no" questions which puts them on a 0-1 scale, while pricepercent and sugarpercent are also on a 0-1 sclae, they are not being calculated on a "yes/no" basis so the values are not just 0 and 1, while for winpercent it is also on an entirely different scale, that being 0-100.

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

A zero would indicate that the candy is not a chocolate candy, and a one would indicate that the candy is a chocolate candy.

> Q8. Plot a histogram of winpercent values

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



> Q9. Is the distribution of winpercent values symmetrical?

No, there is a slight skew to the left.

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

It is below 50%

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

```
# Comparing chocolate and fruity candy
chocolate <- candy$winpercent[as.logical(candy$chocolate)]
fruit <- candy$winpercent[as.logical(candy$fruity)]
```

```
# Finding the mean of chocolate
mean(chocolate)
```

```
[1] 60.92153
```

```
# Finding the mean of fruity candy
mean(fruit)
```

```
[1] 44.11974
```

```
# statistical test
t.test(chocolate, fruit)
```

```

    Welch Two Sample t-test

data:  chocolate and fruit
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 
```

Average Chocolate winpercent: 60.92 Average Fruity Candy winpercent: 44.12

These values prove chocolate is ranked higher with a greater mean winpercent value than that of fruity candy.

> Q12. Is this difference statistically significant?

Yes, this difference is statistically significant as the p-value is 2.81e-08 which is far less than 0.05, the threshold for a statifically signifcant difference. The confidence interval of the difference between the two mean values is also pretty low, which narrows down the data.

3. Overall Candy Rankings

Let's sort the entire dataset by winpercent.

```
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
```

```
tail(candy[order(candy$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
```

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

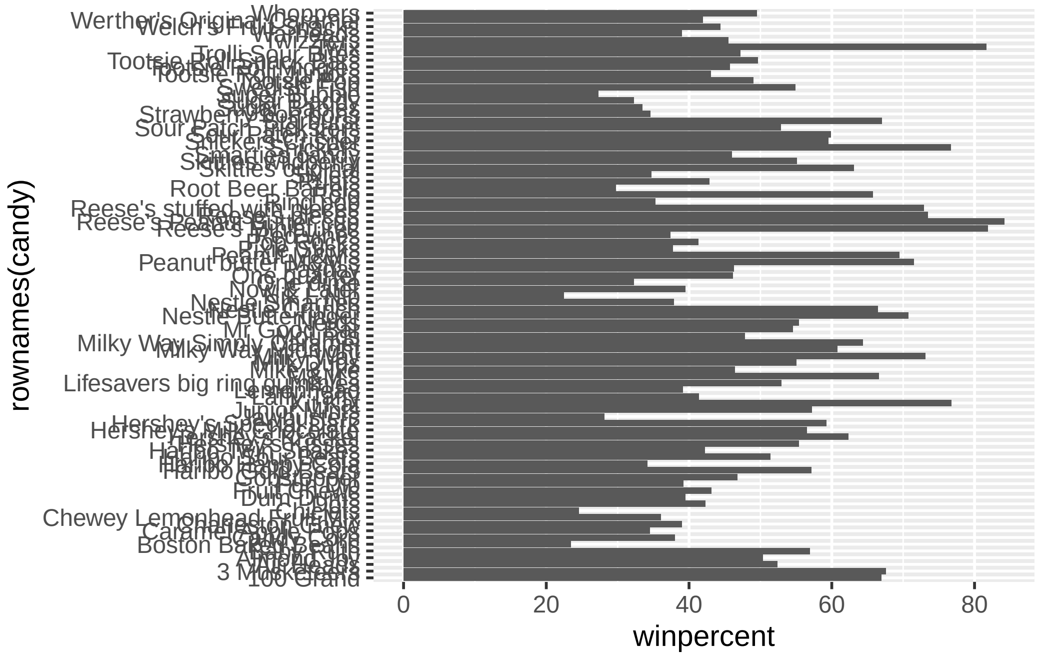
The five least liked candy types in this dataset are: Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

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

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

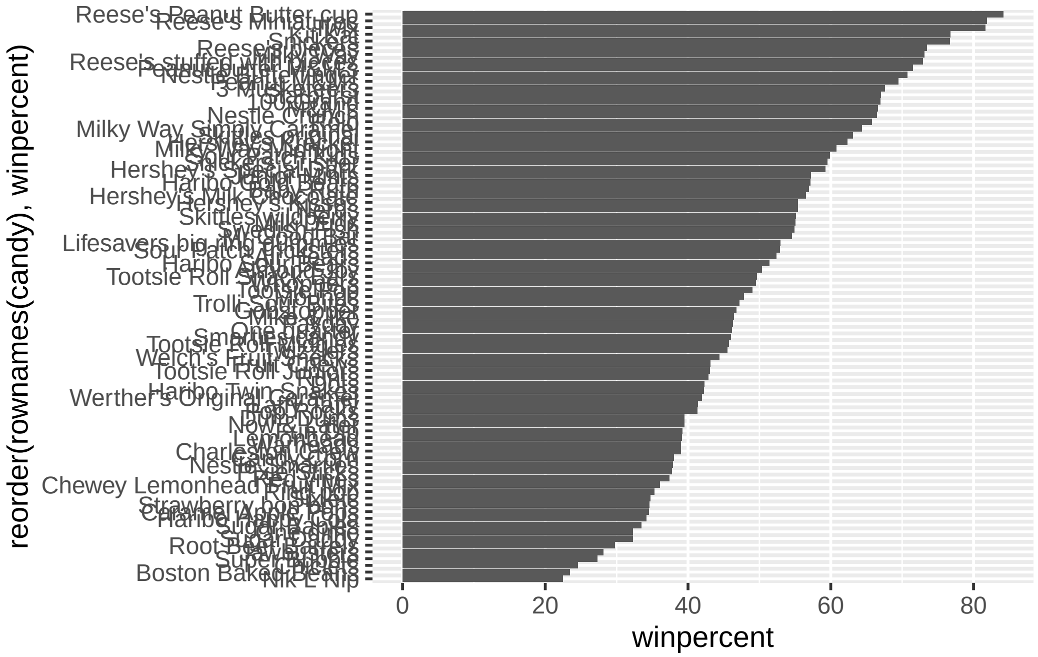
> Q15. Make a first barplot of candy ranking based on winpercent values.

```r
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?

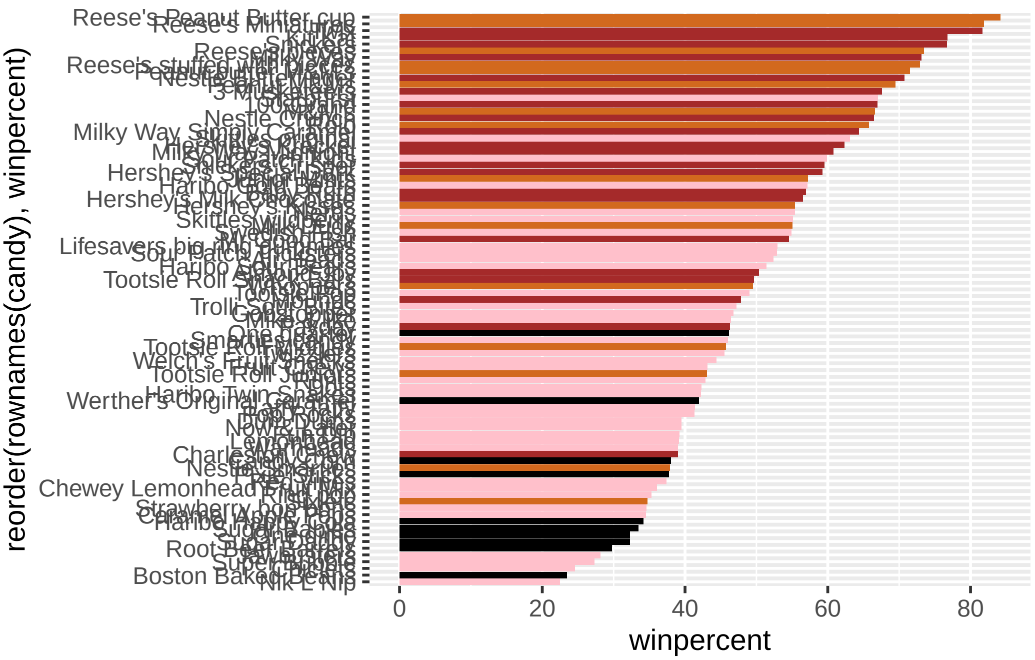
```
library(ggplot2)
ggplot(candy) + aes(winpercent, reorder(rownames(candy),winpercent)) + geom_col()
```



Now let's try adding some color to this plot.

```
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)
```



Q17. What is the worst ranked chocolate candy?

The worst ranked chocolate candy is: Sixlets

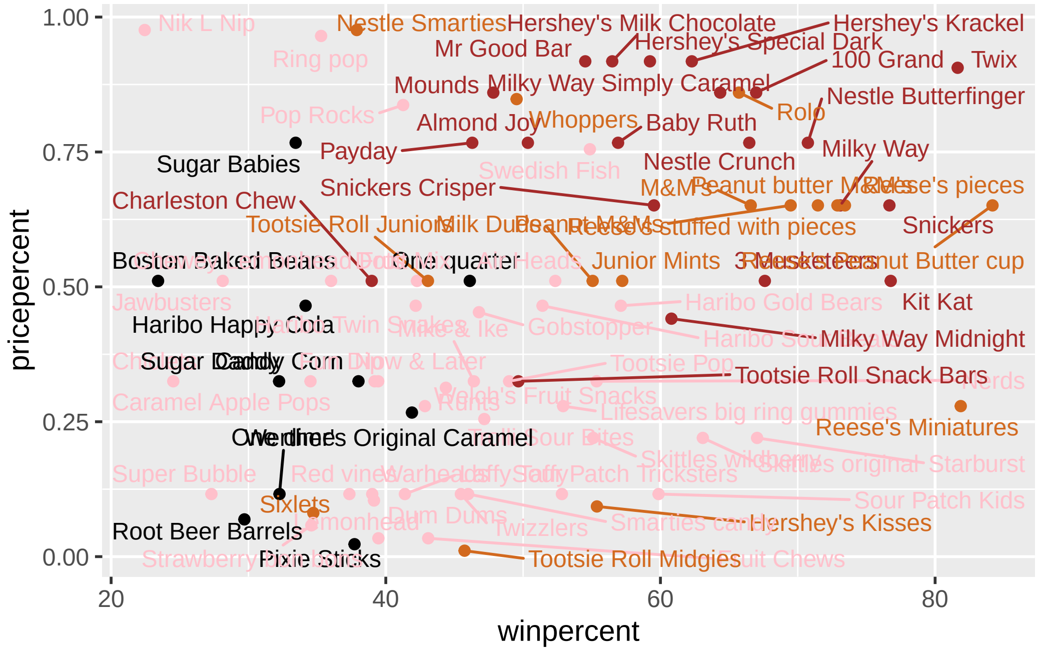
Q18. What is the best ranked fruity candy?

The best ranked fruity candy is: Starburst

4. Taking a look at pricepercent

Now let's now plot our winpercent as pricepercent

```
# install.packages('ggrepel')
library(ggrepel)

options(ggrepel.max.overlaps = Inf)

# 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 = 60)
```



```
# Trying to find most expensive, least popular
ord1 <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord1,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
```

```r
# Trying to find least expensive, most popular
ord2 <- order(candy$winpercent, decreasing = T)
head( candy[ord2,c(11,12)], n=5 )
```

```
                          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
```

> 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 type that offers the most bang for your buck would be: Reese's Miniatures, Kit Kat, Snickers, Reese's Peanut Butter cup, Twix, the absolute best bang for buck being: Reese's Miniatures due to the great ratio between pricepercent and winpercent.

> 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, Ring pop, Nestle Smarties, Hershey's Milk Chocolate, Hershey's Krackel, with the absolute least popular one being: Nik L Nip.

5. Exploring the correlation structure

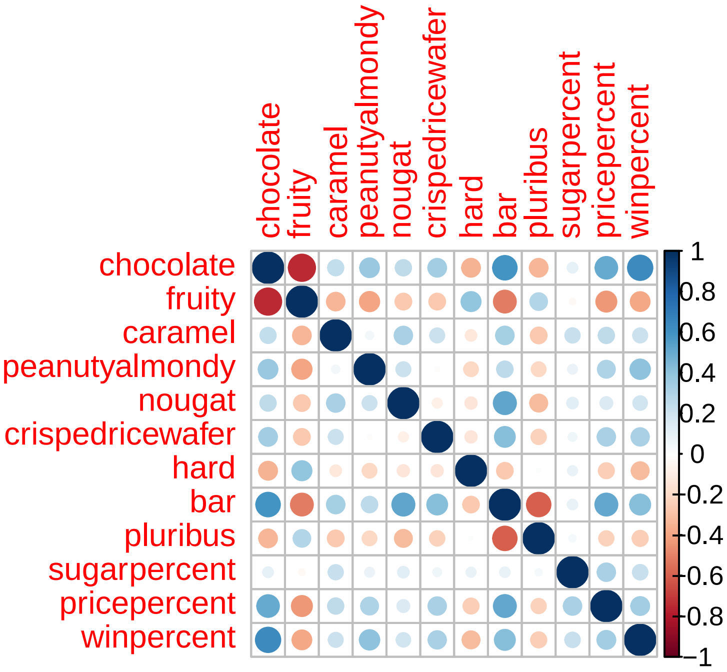
Using corrplot

```r
# install.packages('corrplot')
library(corrplot)
```

```
corrplot 0.92 loaded
```

corrplot 0.92 loaded

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



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

The two variables which are anti-correlated are: 'fruity' and 'chocolate' with nearly a -1 correlation.

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

The two variables that are most positively correlated are: 'bar' and 'chocolate' with a correlation of about 0.8.

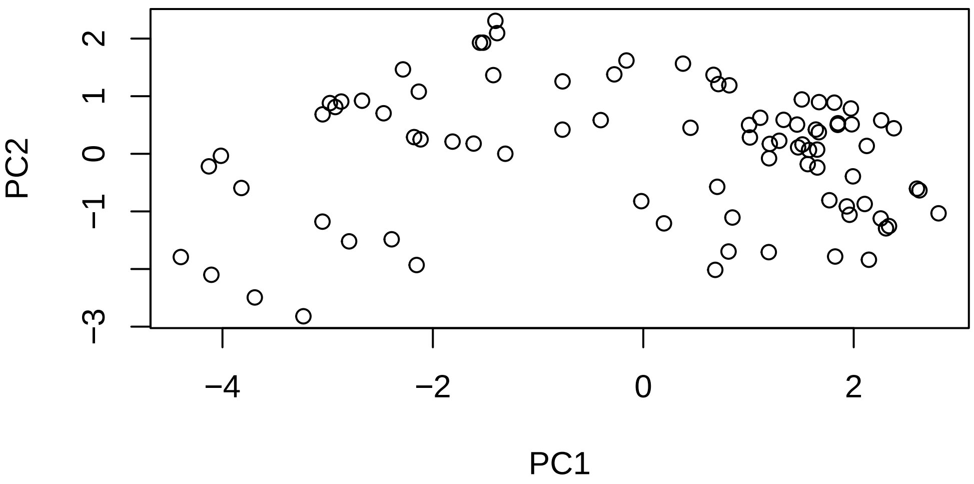
6. Principal Component Analysis

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

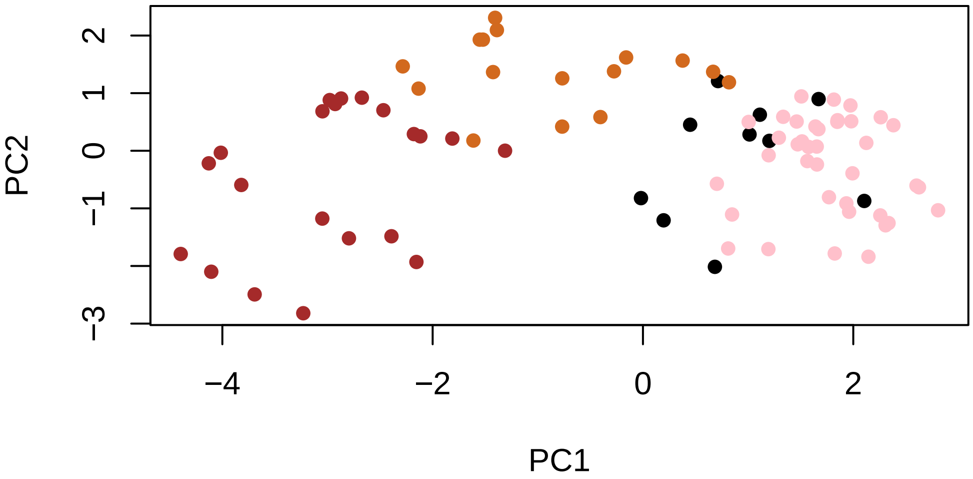
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 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
```

Now we can plot our main PCA score plot of PC1 vs PC2.

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



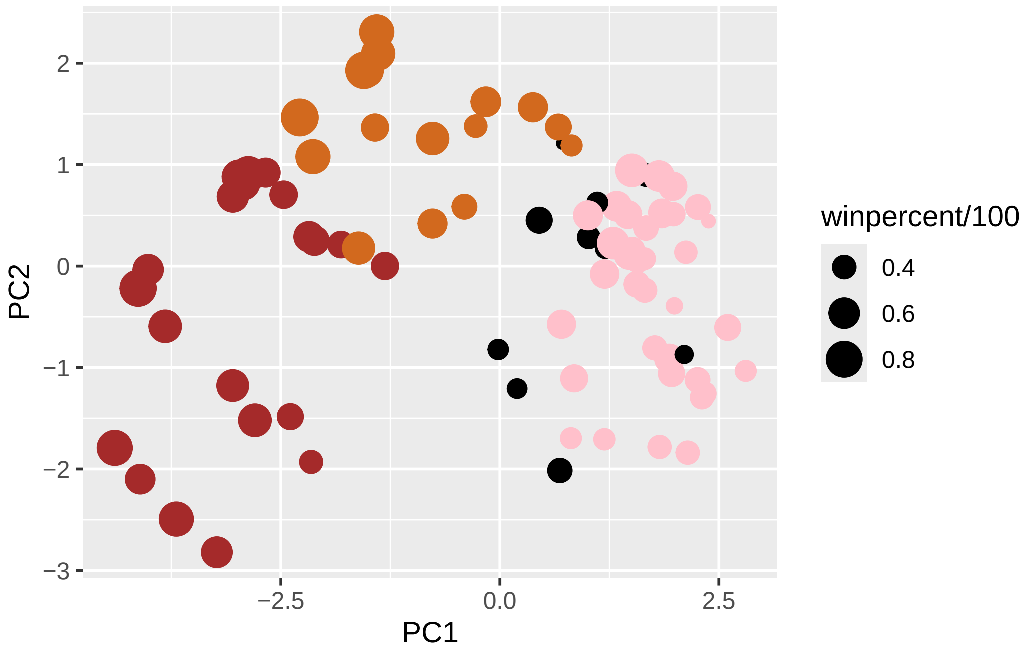
```
# Let's refine the plot
plot(pca$x[,1:2], col=my_cols, pch=16)
```



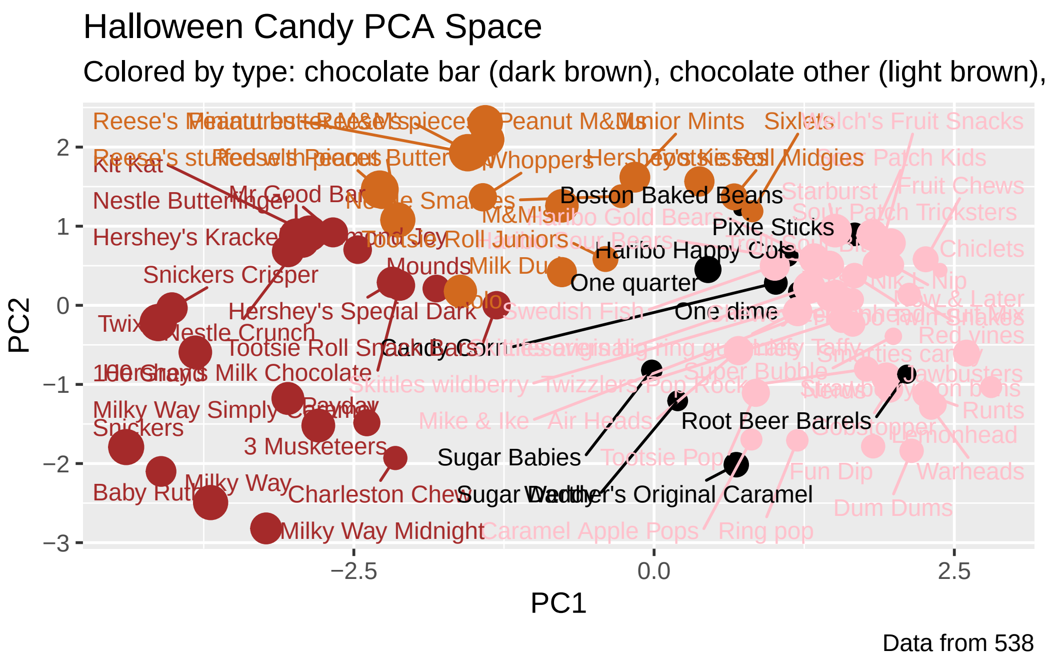
We can make a much nicer plot with the ggplot2 package by making a new data.frame here that contains our PCA results with all the rest of our candy data. We will then use this for making plots below.

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

# plotting the new data frame
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
```



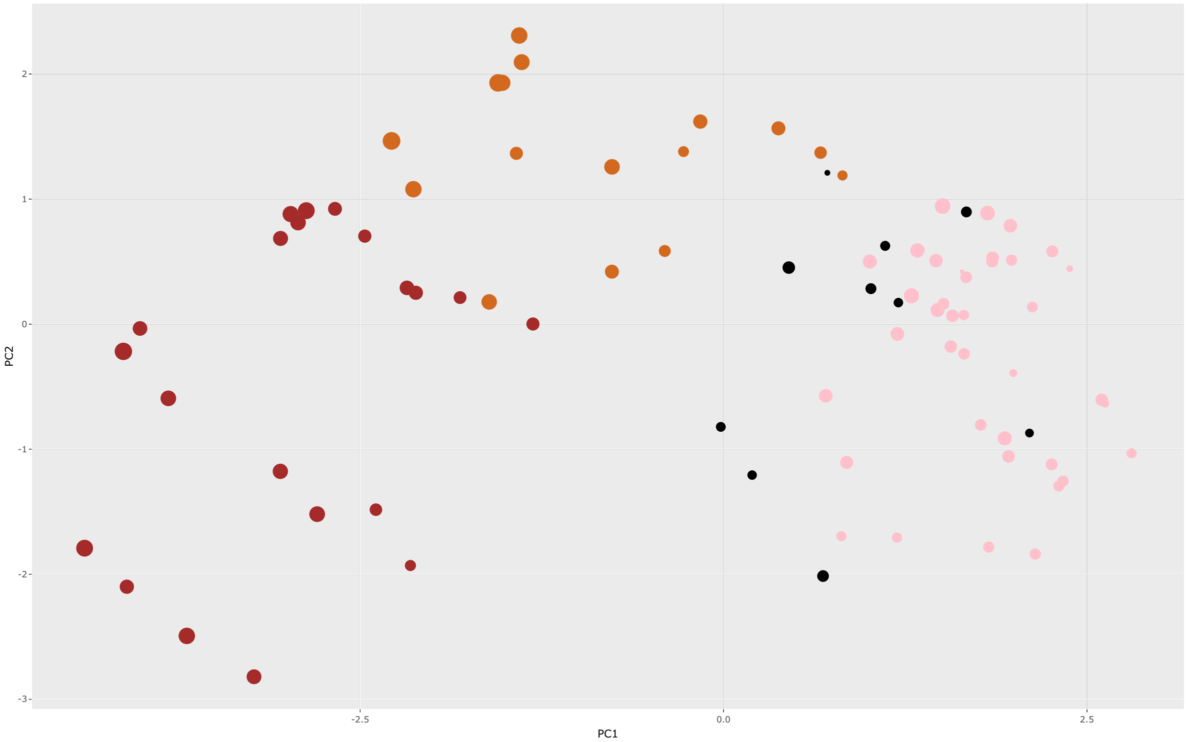
```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 45)  +
  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")
```



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

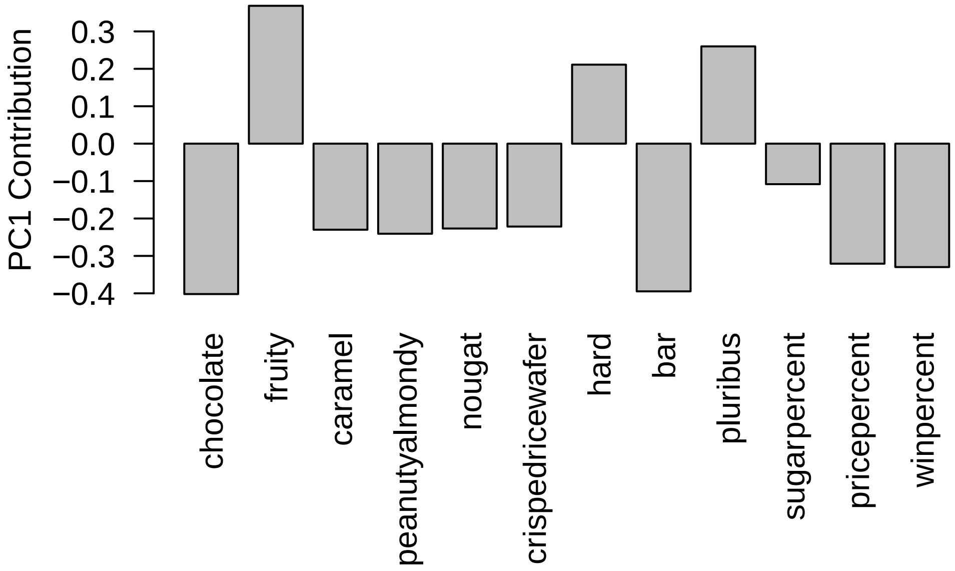
```
Attaching package: 'plotly'

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)
```



Let's take a quick look at our PCA loadings.

```
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 which are picked up strongly by PC1 in the positive direction are: 'fruity', 'hard', and 'pluribus'. These do make sense to me as these variables all are in connection to one another, and it adds up as they have a negative correlation with the chocolate candy related variables.