Class 09: Halloween Mini-Project

Safiya Sayd (PID: A18027139)

Table of contents

1. Importing candy data

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

```
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

##2. What is your favorate candy?

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

```
[1] 81.64291
```

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

My favorite candy is Reese's pieces

```
candy["Reese's pieces", ]$winpercent
```

```
[1] 73.43499
```

Q4. What is the winpercent value for "Kit Kat"?

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

"skimr" package gives you a quick overview of dataset

```
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?

winpercent has a different scale compared to the other variables with a scale of 0-100 while the rest are on a scale of 0-1.

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

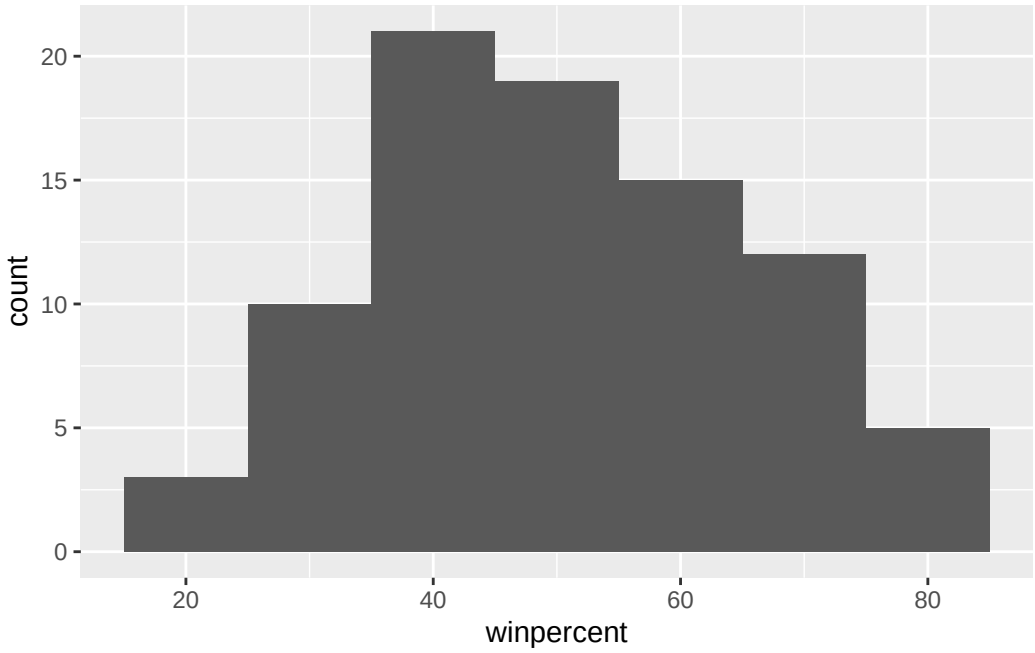
```
candy$chocolate
```

```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

I think it represents if the candy is or isn't chocolate based. It's assigned a value of 1 if its a chocolate based candy and 0 if it isn't.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)
ggplot(candy) +
  aes(winpercent) +
  geom_histogram(binwidth = 10)
```



Q9. Is the distribution of winpercent values symmetrical?

The distribution is not symmetrical, there is a longer tail on the right side. Distribution is skewed to right.

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

To find whether the center of distribution is above or below 50% you can use the median. If the value is less that 50 percent the center is below 50 and if its above 50% the center is above 50 percent.

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

```
[1] 47.82975
```

The center of distribution is below 50%.

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

```
mean(candy$winpercent[candy$chocolate==1])
```

```
[1] 60.92153
```

```
mean(candy$winpercent[candy$fruity==1])
```

```
[1] 44.11974
```

Chocolate is higher ranked then fruity

Q12. Is this difference statistically significant?

t test helps us compare the average win percentage of chocolate compared to fruity

```
t.test(winpercent ~ chocolate, data = candy, var.equal = FALSE)
```

```
    Welch Two Sample t-test

data:  winpercent by chocolate
t = -7.3031, df = 67.539, p-value = 4.164e-10
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -23.91110 -13.64744
sample estimates:
mean in group 0 mean in group 1
       42.14226        60.92153
```

p-value tells you whether the difference is seen is real or by chance. When the p value is greater the 0.05 it safe to say that it is statistically significant but if its equal to or less then 0.05 it isn't significant and considered random!

To test the reliability of the t-test it is important to see if the win percentage is normally distributed using the normality test.

```
shapiro.test(candy$winpercent[candy$chocolate ==1])
```

```
    Shapiro-Wilk normality test

data:  candy$winpercent[candy$chocolate == 1]
W = 0.98088, p-value = 0.7616
```

```
shapiro.test(candy$winpercent[candy$fruity ==1])
```

```
    Shapiro-Wilk normality test

data:  candy$winpercent[candy$fruity == 1]
W = 0.98718, p-value = 0.9342
```

p-value is more the 0.05, the data is normal! This means the t-test can be trusted. It is safe to assume that the difference is not statistically significant.

##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
```

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

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

```
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
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
```

```
library(dplyr)
candy %>% arrange(desc(winpercent)) %>% head(5)
```

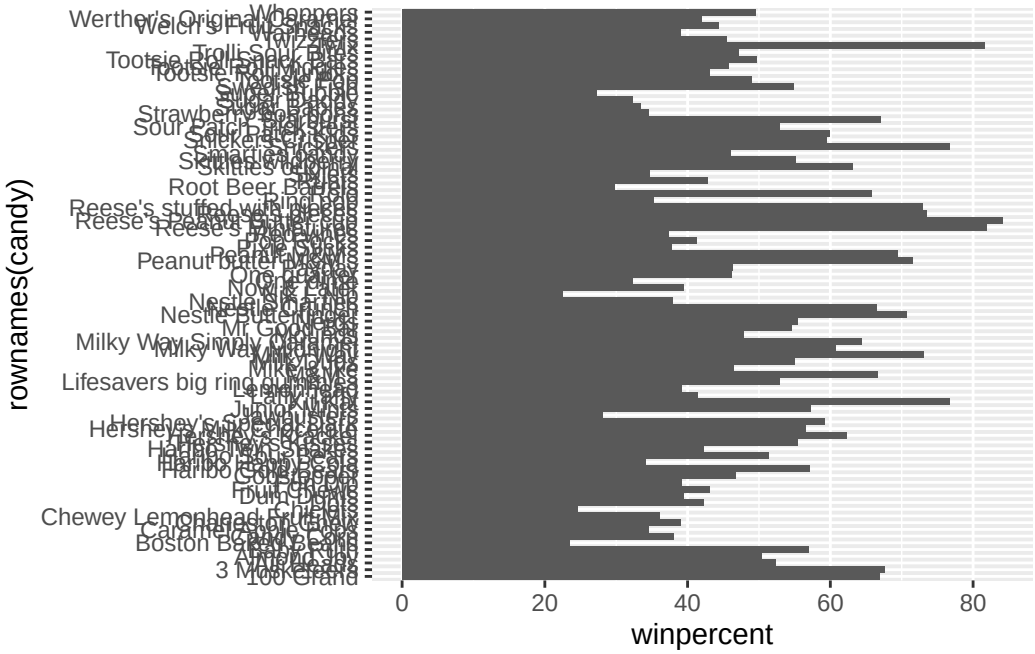
```
                        chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           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
```

I prefer using dplyr because its cleaner and easier to read.

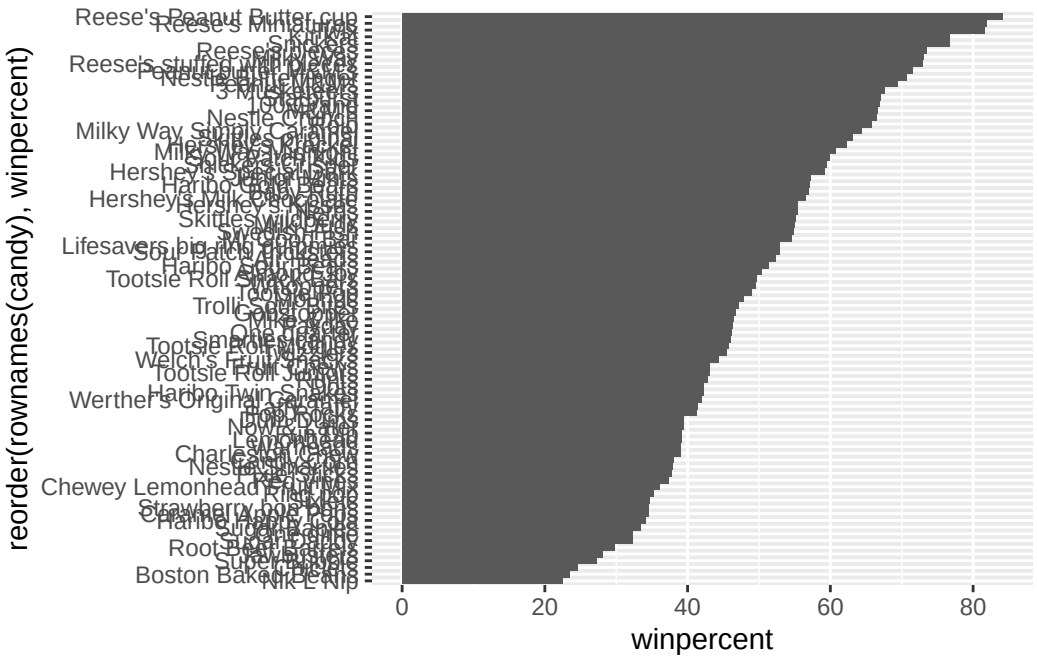
Q15. Make a first barplot of candy ranking based on winpercent values.

```
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?

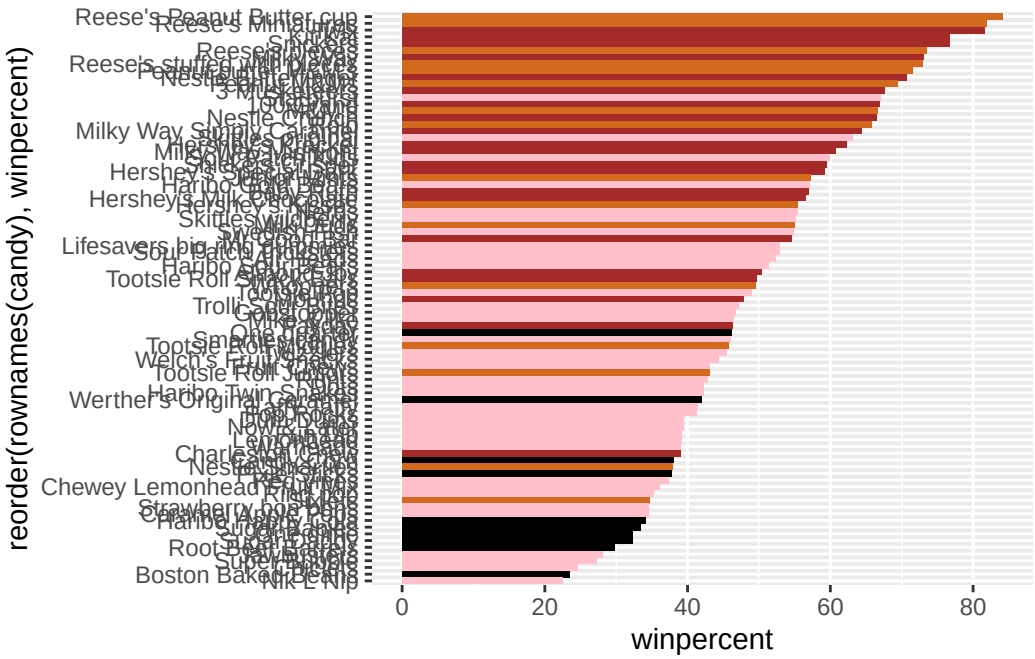
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent) ) +
  geom_col()
```



Customizing a color scheme by candy type

```
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?

Sixlets

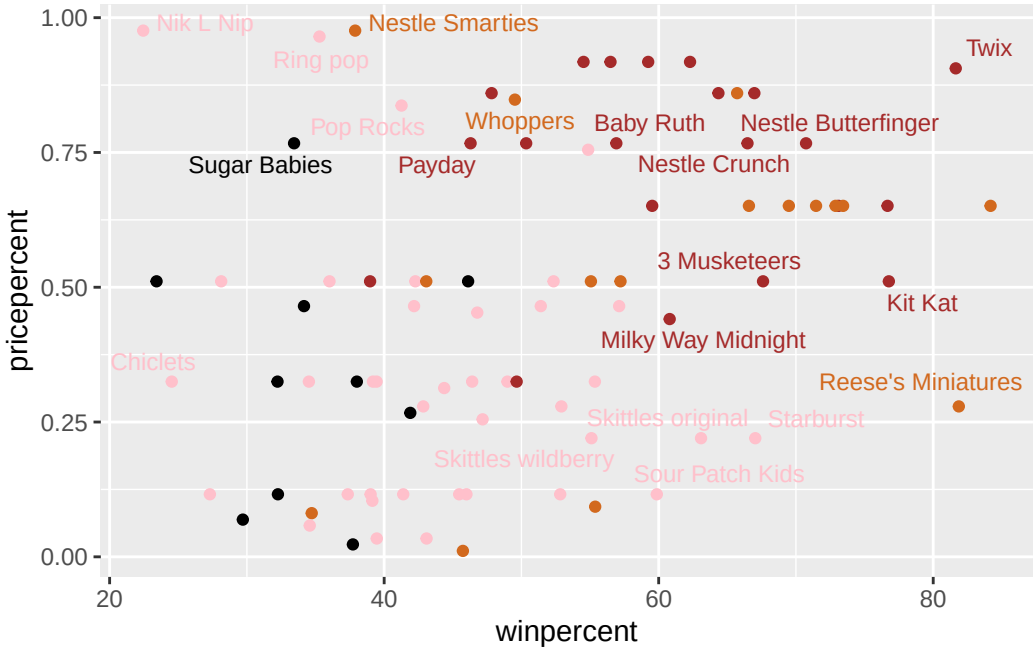
Q18. What is the best ranked fruity candy?

Starbursts

##4. Taking a look at pricepercent

```
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?

Chocolate candies, specifically Reese's miniture

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
```

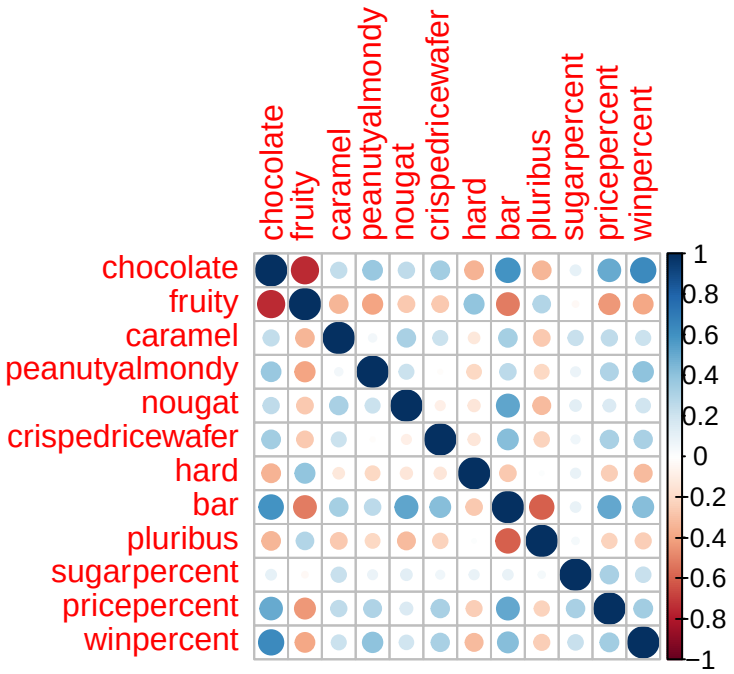
##5 Exploring the correlation structure

seeing how variables interact with one another

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



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

Chocolate and fruity

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

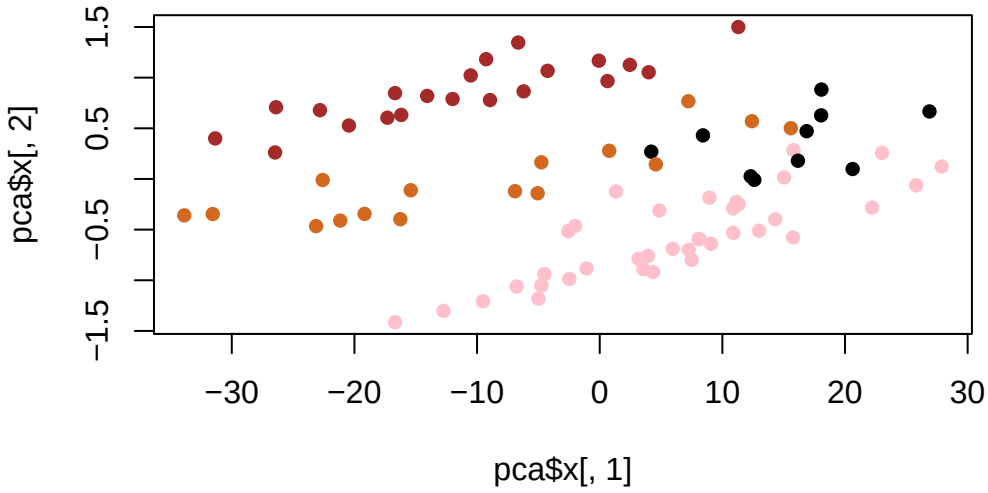
Chocolate and bar

##6. Principal Component Analysis

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

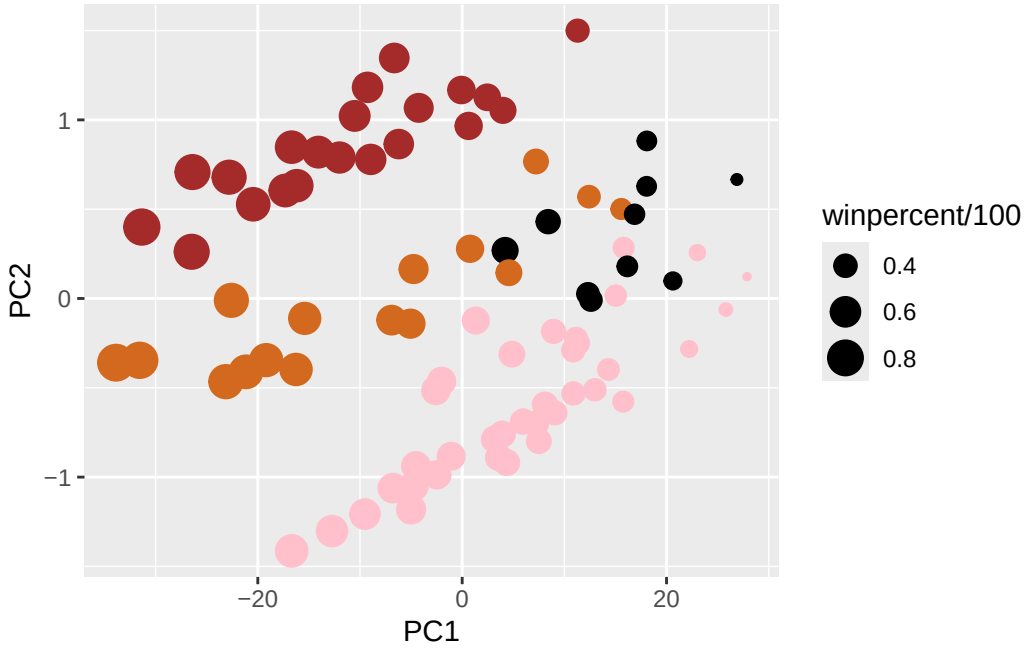
```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     14.7231 0.70241 0.47762 0.37292 0.34641 0.33614 0.30748
Proportion of Variance  0.9935 0.00226 0.00105 0.00064 0.00055 0.00052 0.00043
Cumulative Proportion   0.9935 0.99574 0.99678 0.99742 0.99797 0.99849 0.99892
                          PC8     PC9    PC10    PC11    PC12
Standard deviation     0.27417 0.23826 0.21435 0.18434 0.15331
Proportion of Variance 0.00034 0.00026 0.00021 0.00016 0.00011
Cumulative Proportion  0.99927 0.99953 0.99974 0.99989 1.00000
```

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



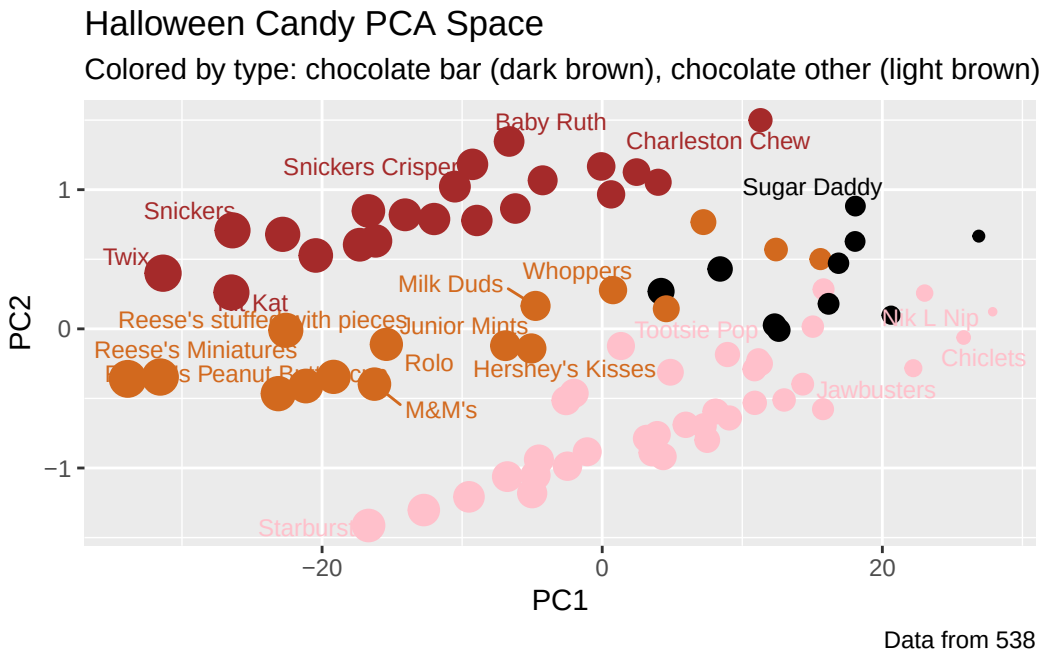

```r
my_data <- cbind(candy, pca$x[,1:3])
```

```r
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 brown),
       caption="Data from 538")
```

```
Warning: ggrepel: 64 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

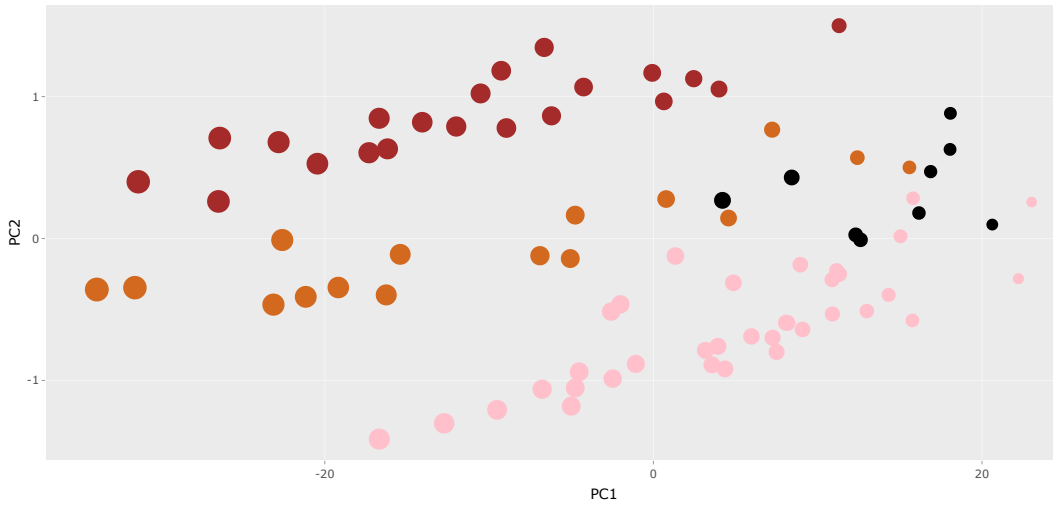



```
library(plotly)
```

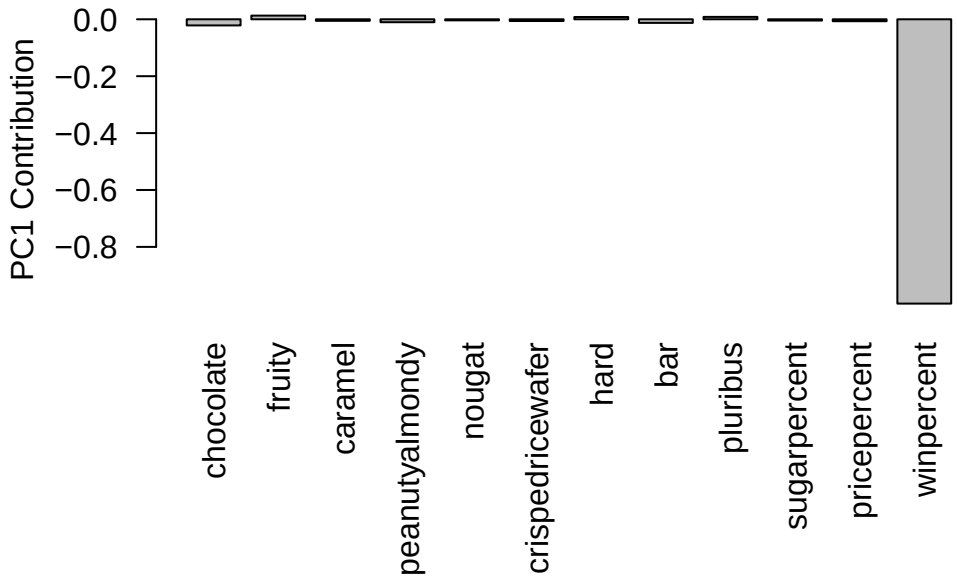
```
Attaching package: 'plotly'

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/sy/_fr9v5r51nxc7bzp4h_683nh0000gn/T/RtmpLaan3L/file6fdc4b857c51,
```


PC2
1
0
-1
-20
0
20
PC1

```
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 strongly by PC1 in positive direction. This makes sense to me.