Class 10: Candy Mini Project

Michelle Woo

1. Importing the data

```
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 data set?

```
total_candy_types <- nrow(candy)

total_candy_types
```

```
[1] 85
```

85 different candy types

Q2. How many fruity candy types are in the data set?

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

fruit_candy_types
```

```
[1] 38
```

38 fruity candy types

2. What is your favorite candy?

using `winpercent` to determine the more popular candy

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

```
[1] 81.64291
```

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

```
candy["Sour Patch Kids", ]$winpercent
```

```
[1] 59.864
```

My favorite candy is Sour Patch Kids and its `winpercent` value is around 59.86

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

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

```
[1] 76.7686
```

Around 76.77

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

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

```
[1] 49.6535
```

Around 49.65, it is not as popular as Kit Kat according to their `winpercent` value.

Getting 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	
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?

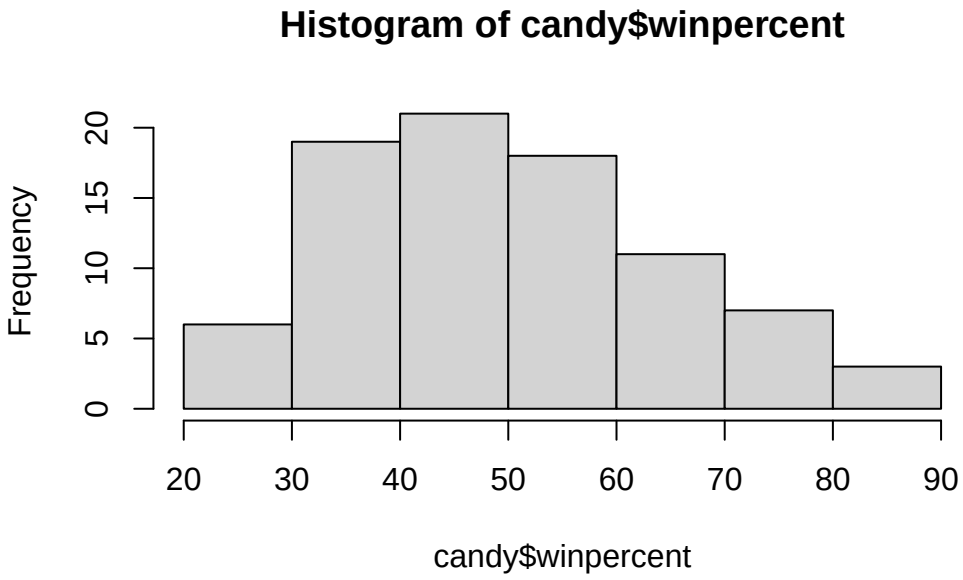
Most of the variables fall under a yes or no question , such as if it is a chocolate or fruity type of candy. That is reflected by the values going from 0 to 1. The variables that are on a different scale would be winpercent, pricepercent, and sugarpercent. Those are calculated from 0 to 100.

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

A zero would represent that the candy is not chocolaty and one represents that it is a chocolate candy.

Q8. Plotting the data as a histogram:

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



Q9. Is the distribution of `winpercent` values symmetrical?

No, they are not symmetrical. The distribution is slightly skewed to the left.

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

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
mean(chocolate)
```

```
[1] 60.92153
```

```
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
```

Chocolate: 60.92

Fruity: 44.12

On average, chocolate is ranked higher at 60.92 than fruity candy which has mean of 44.12.

Q12. Is this difference statistically significant?

The difference is statistically significant as the p value is much below 0.05. The confidence interval of the difference between the means is also quite low which narrows down the data.

3. Overall Candy Rankings

Sorting the whole 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?

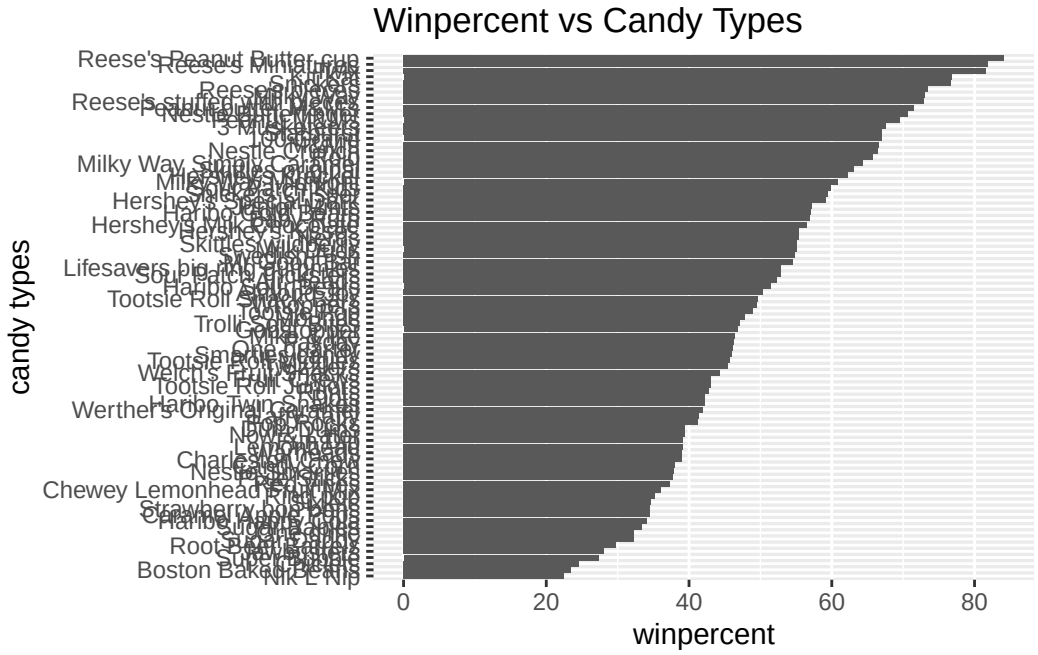
The five least liked candy types are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

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

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

Plotting the data using ggplot:

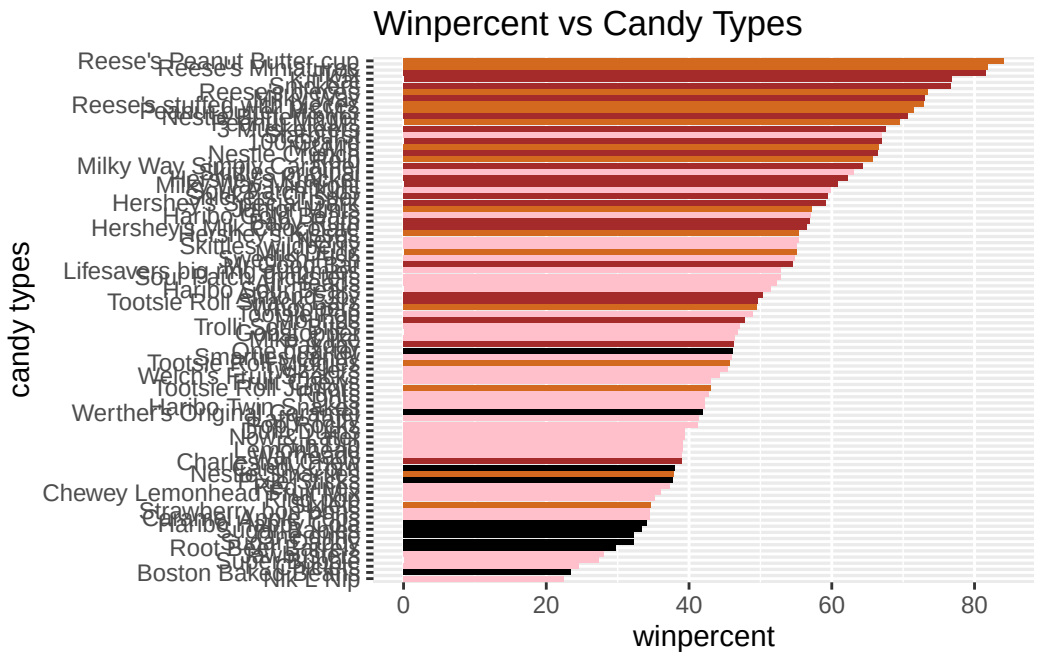
```
library(ggplot2)

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



Adding color to the 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_
```



Q17. What is the worst ranked chocolate candy?

Sixlets

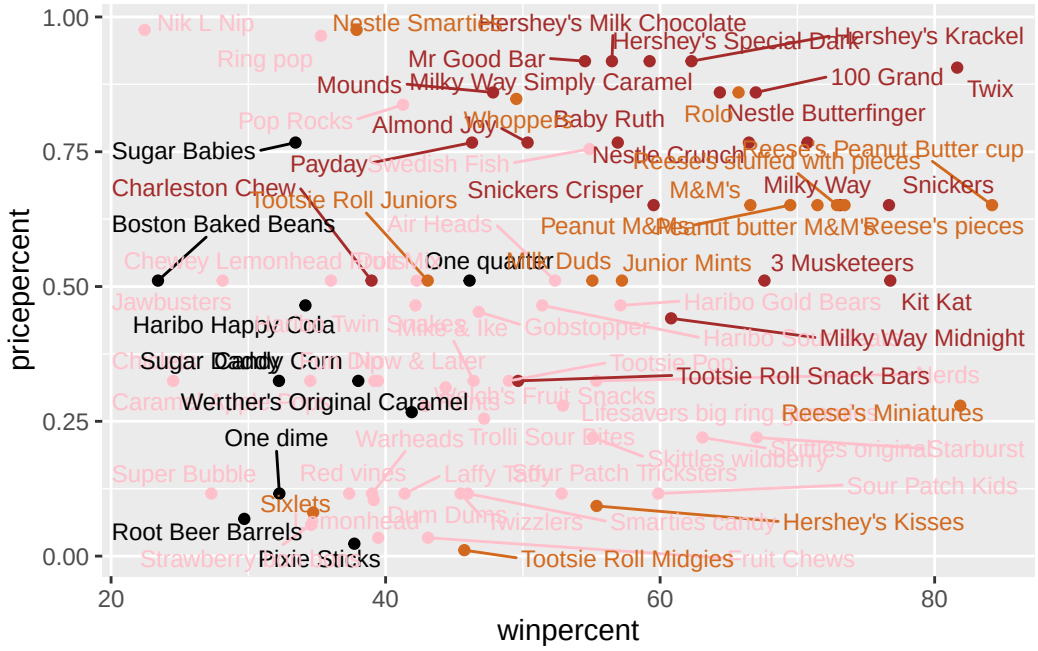
Q18. What is the best ranked fruity candy? Starburst

4. Looking at the price point

Plotting the `winpercent` vs `pricepercent`

```
library(ggrepel)
```

```
# plotting 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)
```




```
# 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
```

```
# 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?

Reese's Miniatures, Kit Kat, Snickers, Reese's Peanut Butter cup, Twix

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

Nik L Nip, Ring pop, Nestle Smarties, Hershey's Milk Chocolate, Hershey's Krackel

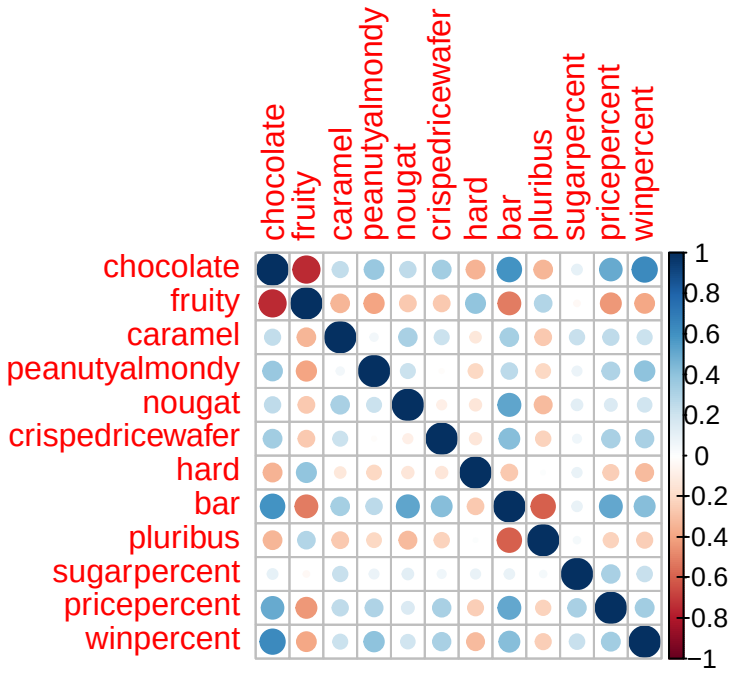
5. Exploring the correlation structure

using `corrplot`

```
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 and chocolate have a nearly -1 correlation.

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

Bar and chocolate are around 0.8 which means they are the most positively correlated.

6. PCA

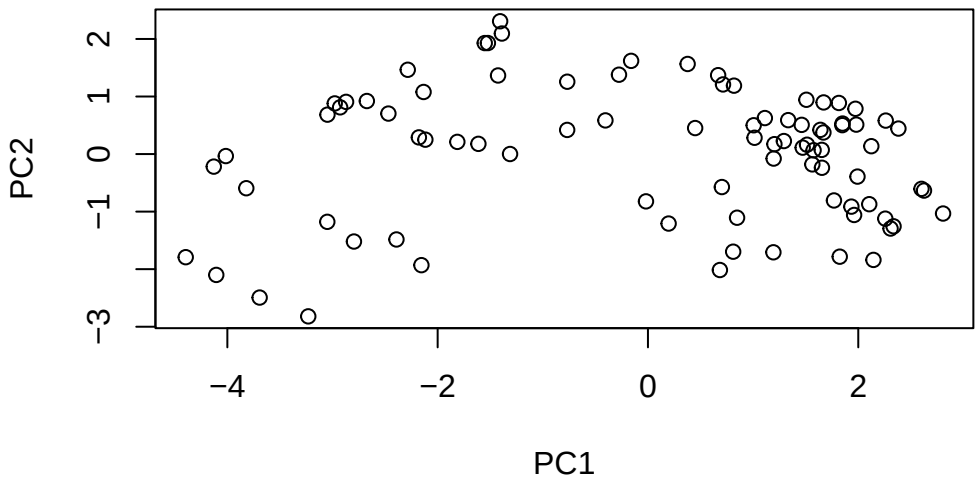
```
pca <- prcomp(candy, scale = T)

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
```

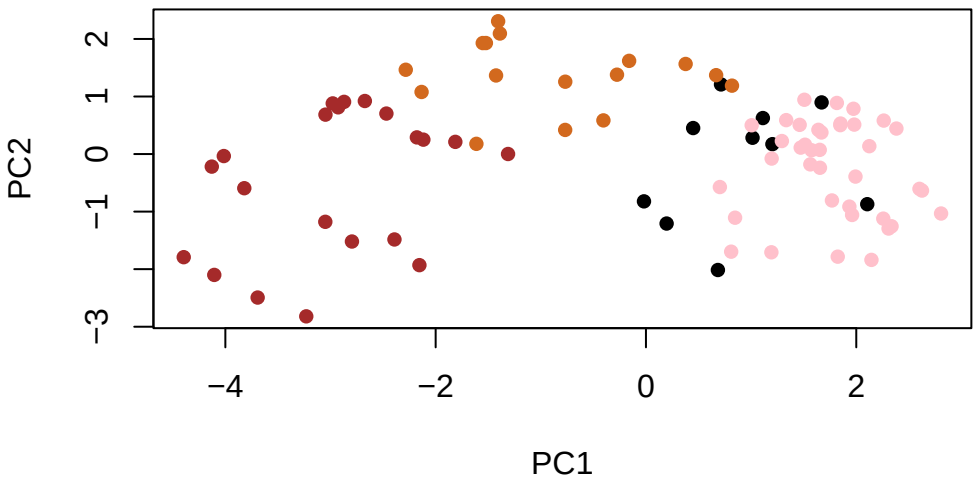
```
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

Plotting the main PCA score of PC1 vs PC2:

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

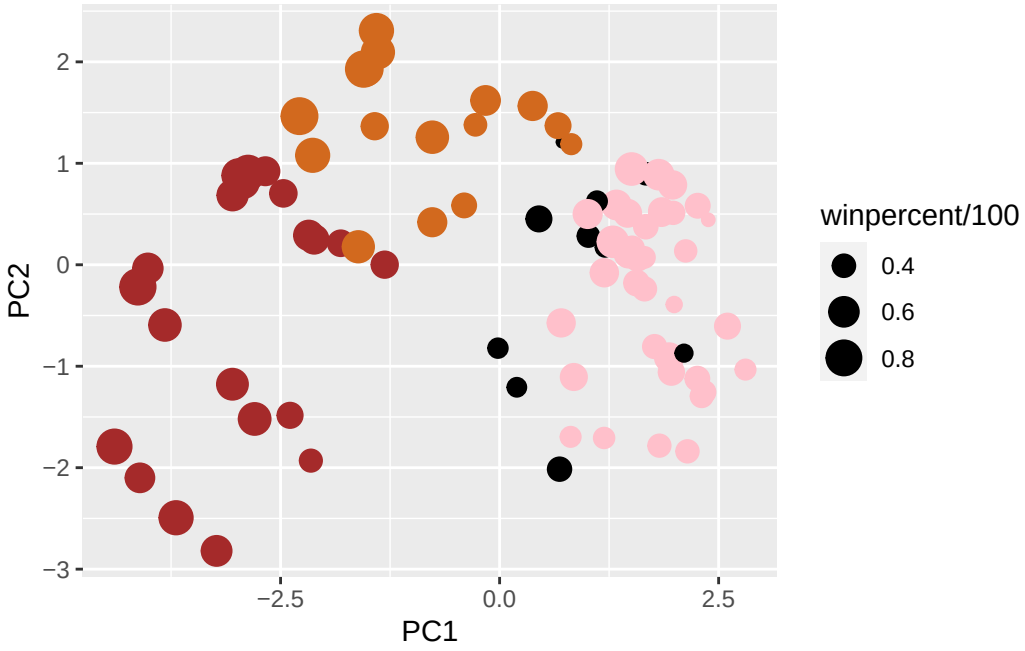



```
# refining the plot
plot(pca$x[,1:2], col=my_cols, pch=16)
```

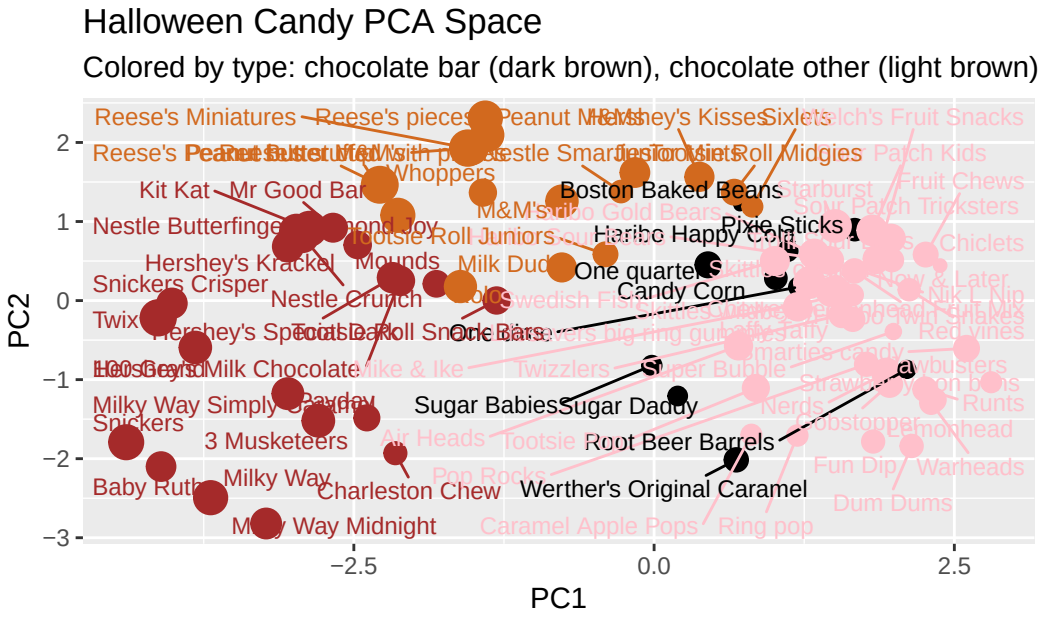


Now we will be making a new data frame that has the PCA results with all of the candy data. This will make ggplot work best. The new data frame should include a separate column for each of the aesthetics displayed in the final plot.

```
# 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
```




```
# touching up the plot
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 37)  +
  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")
```

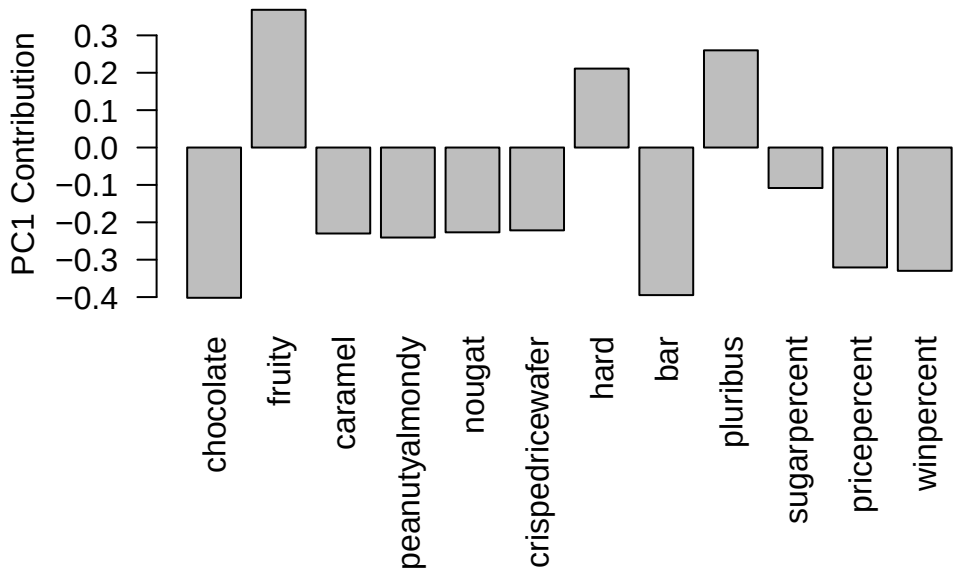


Data from 538

```
# making the plot easier to read
# install.packages(plotly)

# library(plotly)

# ggplotly(p)
```

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?**

Fruity, hard, and pluribus are the variables that are picked up strongly by PC1 in the positive direction. This does make sense, these variables would correlate together and it would be negatively correlated to the chocolate related variables.