Class 11

Raidah Anisah Huda

Candy data

In today's class we will examine S3* Candy data and see if this helps us gain some more feeling for how PCA and other methods work.

```
candy <- read.csv("candy-data.csv", 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
```

There are 85 different candy types

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

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

```
[1] 38
```

There are 38 fruity candy types.

Q. What are these fruity candy?

We can use the ==

```
row.names(candy[candy$fruity == 1, ])
```

```
 [1] "Air Heads"                    "Caramel Apple Pops"
 [3] "Chewey Lemonhead Fruit Mix"   "Chiclets"
 [5] "Dots"                         "Dum Dums"
 [7] "Fruit Chews"                  "Fun Dip"
 [9] "Gobstopper"                   "Haribo Gold Bears"
[11] "Haribo Sour Bears"            "Haribo Twin Snakes"
[13] "Jawbusters"                   "Laffy Taffy"
[15] "Lemonhead"                    "Lifesavers big ring gummies"
[17] "Mike & Ike"                   "Nerds"
[19] "Nik L Nip"                    "Now & Later"
[21] "Pop Rocks"                    "Red vines"
[23] "Ring pop"                     "Runts"
[25] "Skittles original"            "Skittles wildberry"
[27] "Smarties candy"               "Sour Patch Kids"
[29] "Sour Patch Tricksters"        "Starburst"
[31] "Strawberry bon bons"          "Super Bubble"
[33] "Swedish Fish"                 "Tootsie Pop"
[35] "Trolli Sour Bites"            "Twizzlers"
[37] "Warheads"                     "Welch's Fruit Snacks"
```

How often does my favorite candy win

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

```
[1] 81.64291
```

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

```
candy["Swedish Fish", ]$winpercent
```

```
[1] 54.86111
```

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::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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

```
library("skimr")
```

Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

Yes, a majority of the columns are in a 0:1 scale but the `winpercent` column is on a different scal to the majoirty, with a scale of 0:100.

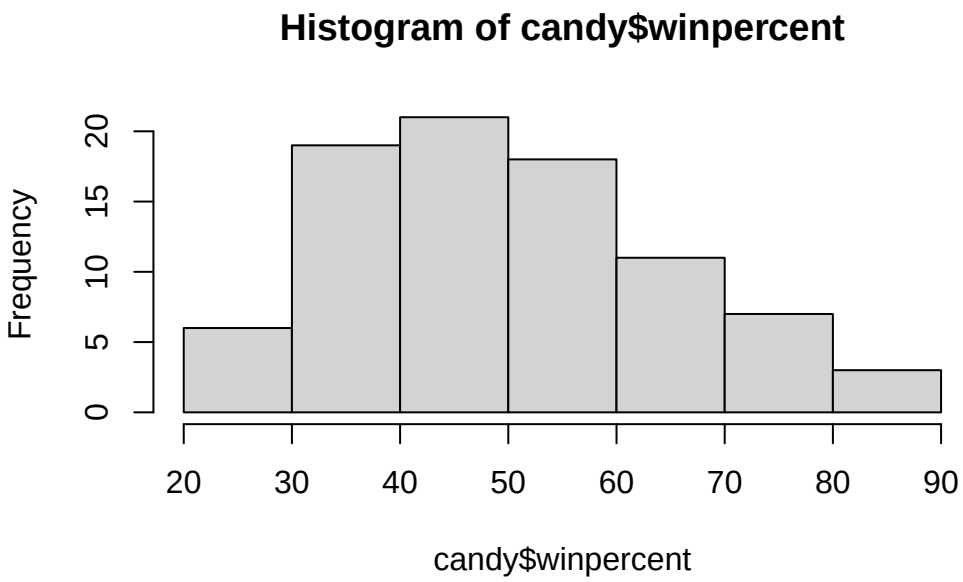
Q7. What do you think a zero and one represent for the candy$chocolate column?

A zero means the candy is not classified as containing chocolate.

Q8. Plot a histogram of winpercent values

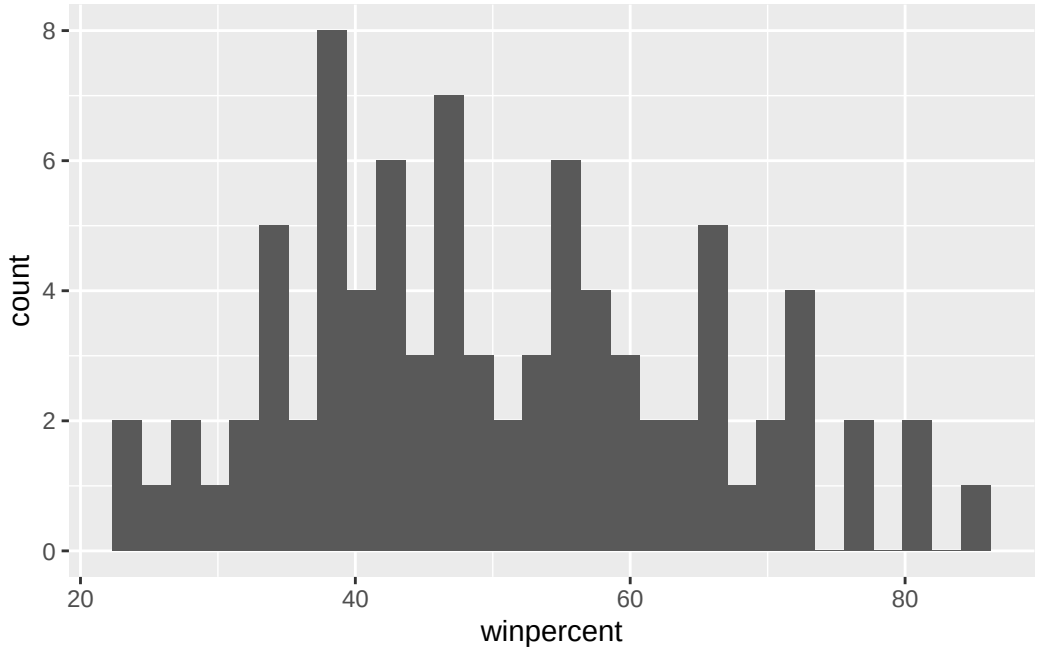
In R basics plot

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




```
library(ggplot2)
hist<- ggplot(candy)+
  aes(winpercent)+
  geom_histogram()

hist
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Q9. Is the distribution of winpercent values symmetrical?

No, it appears that the winpercent values are skewed to the left had side of the histogram, indicating a skew towards lower winpercents.

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

It appears tht the center of the distribution is below 50% with a mean of :

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

```
[1] 50.31676
```

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

To answer this question I will need to: -"subset" (a.k.a. "select", "filter") the candy dataset to just chocolate candy -get there winpercents -calculate the mean of these. Then do the same for fruity candy an compare.

```
#Filter/select/subset to just chocolate rows
choc.candy<- candy[as.logical(candy$chocolate),]
fruity.candy<- candy[as.logical(candy$fruity), ]
```

```
#Get there winpercent
choc.win <- choc.candy$winpercent
fruity.win <- fruity.candy$winpercent

#Calculate their mean winpercent value
mean(choc.win)
```

```
[1] 60.92153
```

```
mean(fruity.win)
```

```
[1] 44.11974
```

On average, the chocolate candy is higher ranked than the fruity candy.

Q12. Is this difference statistically significant?

```
t.test(choc.win, fruity.win)
```

```
    Welch Two Sample t-test

data:  choc.win and fruity.win
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
```

Yes, due to p-value being less than 5%, this difference between chocolate nad fruity candy is statistically significant.

Overall Candy Rankings

There is a base R function called `sort()` for, guess what, sorting vecotrs of input.

```
x<- c(5,2,10)

sort(x, decreasing = TRUE)
```

```
[1] 10  5  2
```

The buddy function to `sort()` that is often even more useful is called `order()`. It returns the "indices" of the input that would result in it being sorted.

```
order(x)
```

```
[1] 2 1 3
```

```
x[order(x)]
```

```
[1]  2  5 10
```

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

I can order by the `winpercent`

```
ord<- order(candy$winpercent)
head(candy[ord,],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
```

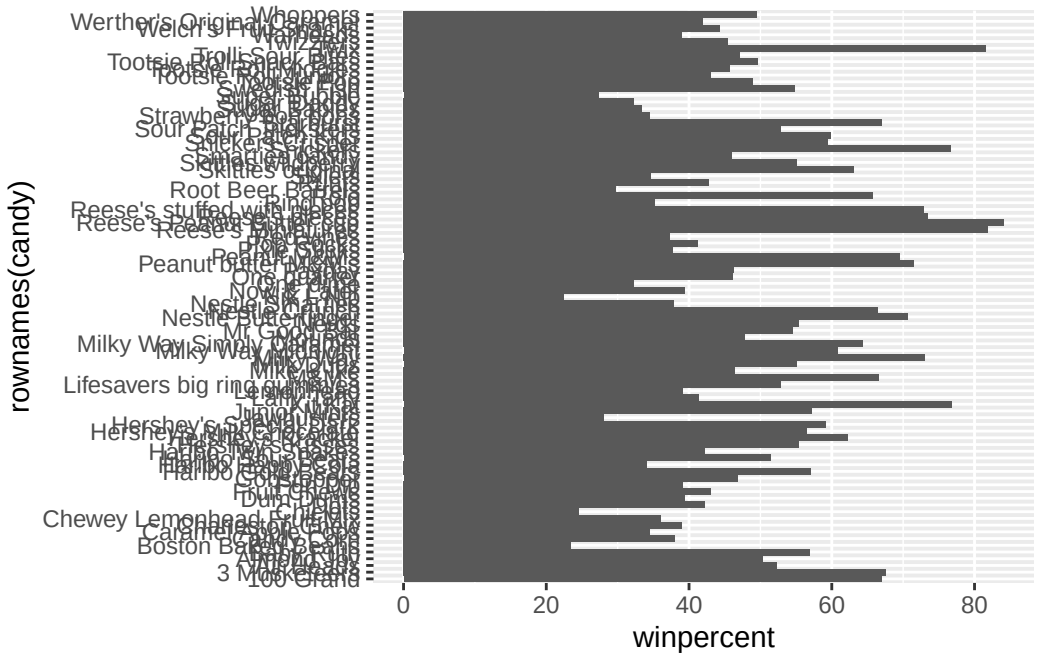
Q14. What are the top 5 all time favorite candy types out of this set?

```
ord.top<- order(candy$winpercent, decreasing = TRUE)
head(candy[ord.top,],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
```

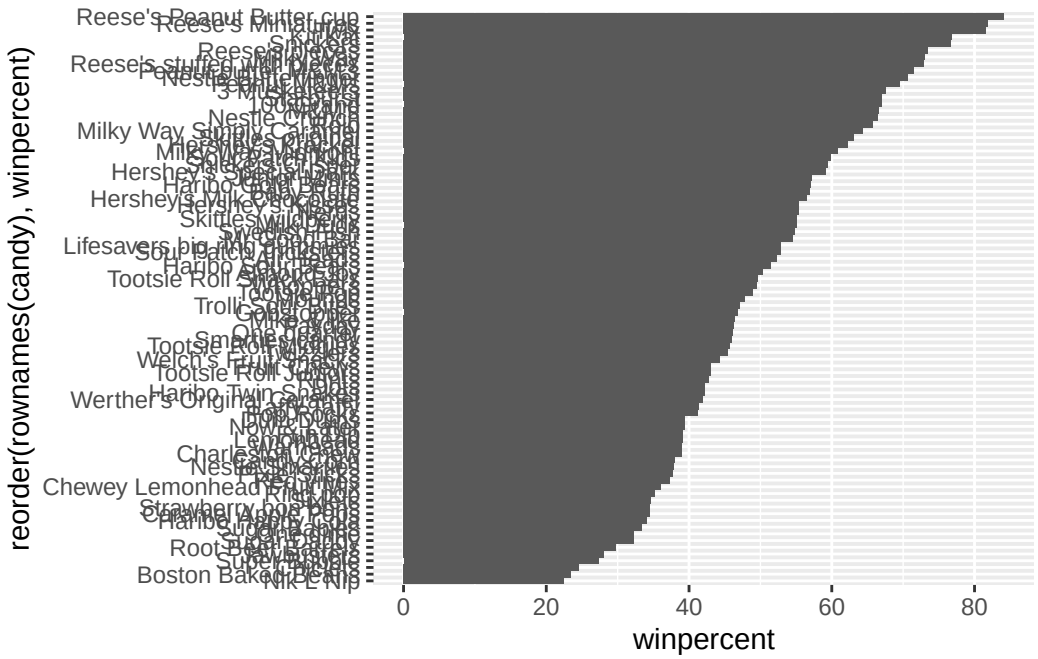
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

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

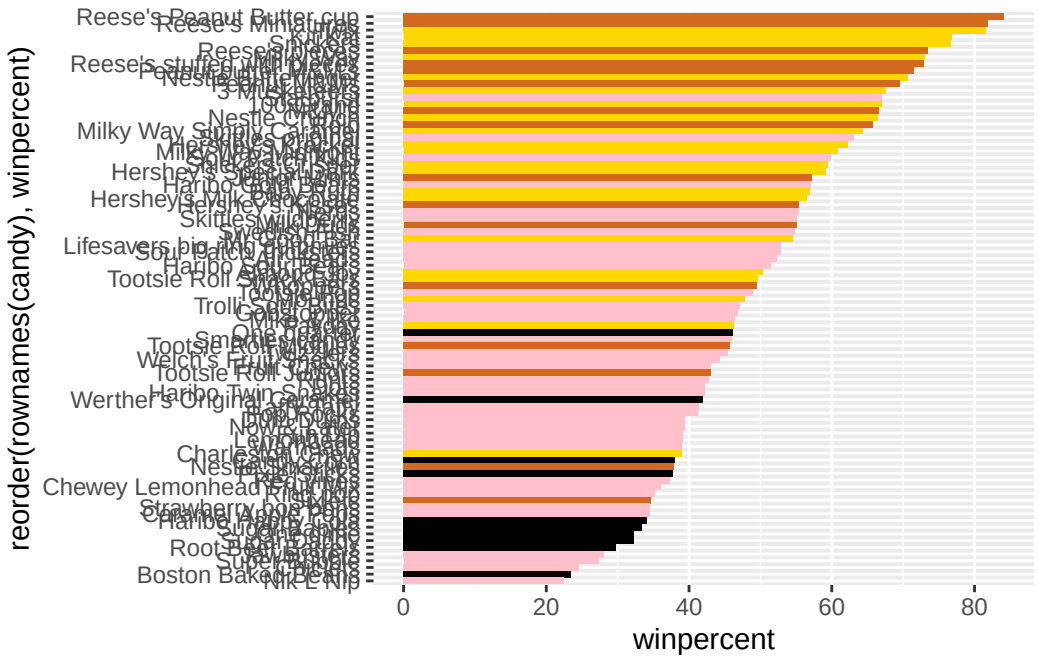


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
library(ggplot2)

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




```
library(ggplot2)

my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "gold"
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?

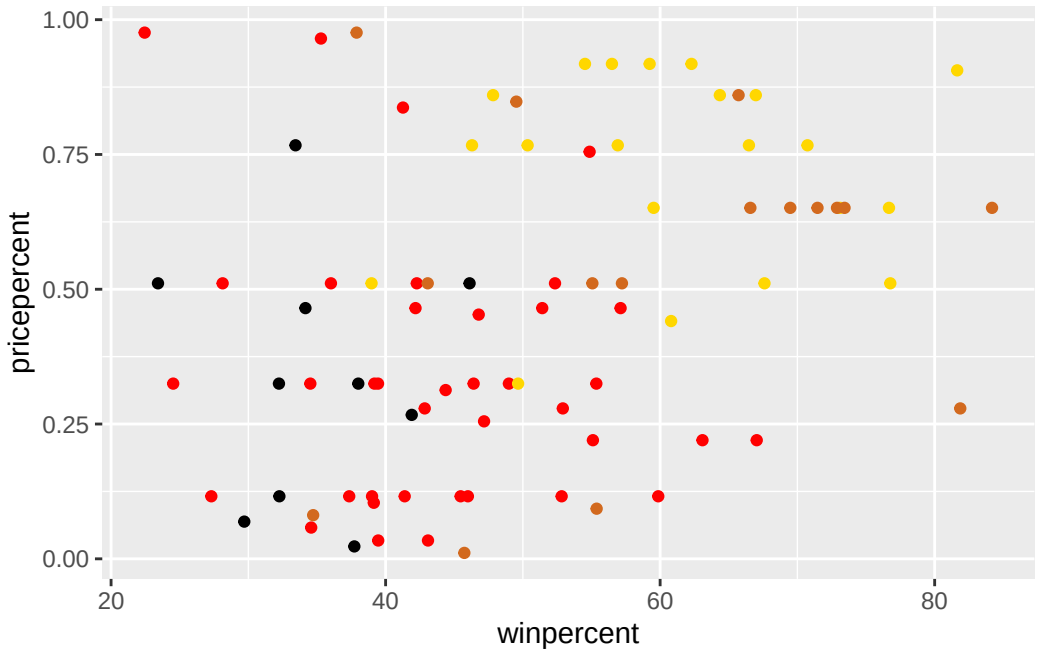
Sixlets are the worst ranked chocolate candy

Q18. What is the best ranked fruity candy?

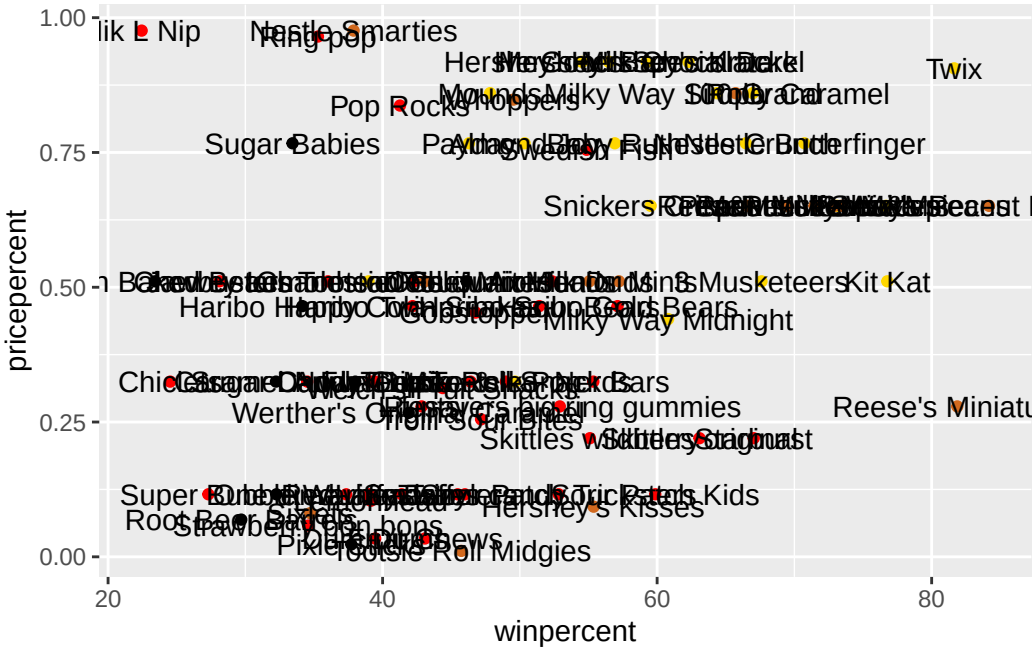
Starburst are the best ranked fruity candy.

Taking a look at Price Percent

```
my_cols[as.logical(candy$fruity)] = "red"
library(ggplot2)

# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent) +
  geom_point(col=my_cols)
```



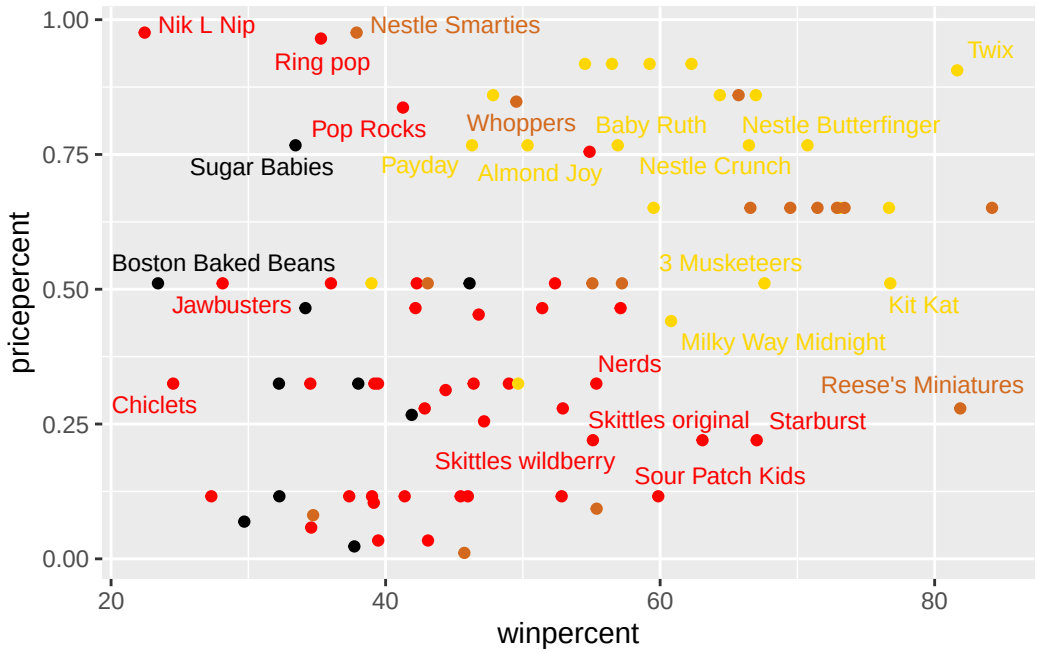

```
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text()
```



To deal with overlapping labels I can use the **geom_repel** package

```
library(ggrepel)

ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(max.overlaps = 6, col=my_cols, size=3)
```

```
Warning: ggrepel: 61 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



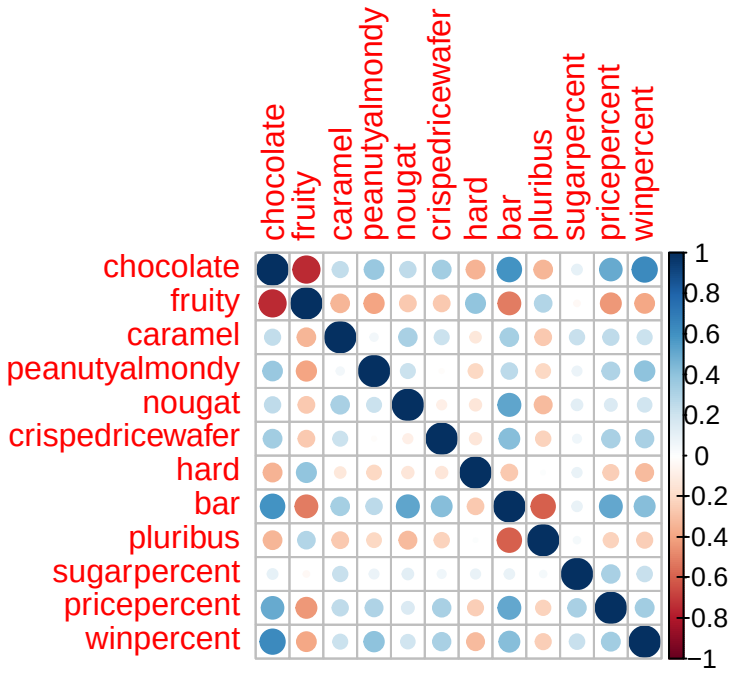
Correlation Structure

Pearson correlation goes between -1 and +1 with zero indicating no correlation and values close to one being very highly correlated

```
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 are anti-correlated

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

Chocolate and Bars appear to be highly correlated as well as chocolate and winpercent

PCA

The base R function for PCA is called `prcomp()` and we can set "scale=TRUE/FALSE"

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

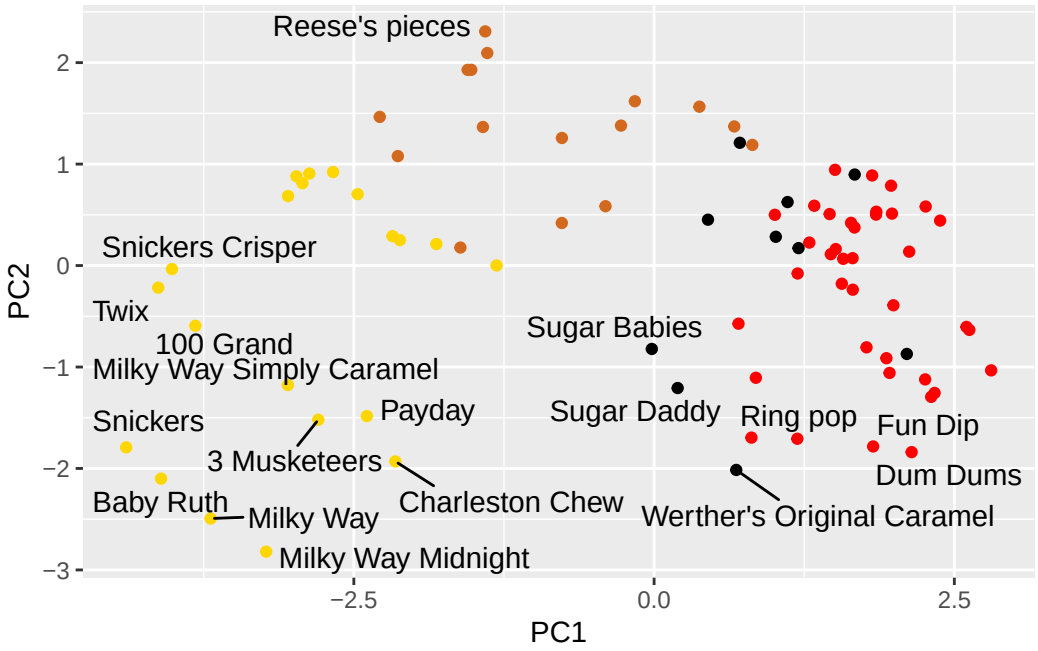
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688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

The main result fo PCA - i.e. the new PC plot (projection of candy on our new PC axis in cotained in 'pca$x

```
pc<- as.data.frame(pca$x)
ggplot(pc)+
  aes(PC1, PC2, label=rownames(pc))+
  geom_point(col=my_cols)+
  geom_text_repel(max.overlaps = 6)
```

```
Warning: ggrepel: 67 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



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 by PC1 in the positive direction and this makes sense.