lab10

Zijing

```
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

```
nrow(candy)
```

```
[1] 85
```

There are 85 candy types in this dataset.

Q2

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

```
[1] 38
```

There are 38 fruity candy types in this dataset.

Q3

```
candy["HersheyÕs Kisses", ]$winpercent
```

```
[1] 55.37545
```

My favorite candy is HersheyÕs Kisses and its winpercent value is 55.37545.

Q4

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

```
[1] 76.7686
```

The winpercent for Kit Kat is 76.7686.

Q5

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

```
[1] 49.6535
```

Tootsie Roll Snack Bars' winpercent is 49.6535.

```
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

winpercent is on a different scale compared to all the other variables, which are all from 0 to 1. Sugarprecent and pricepercent are also relatively different from others as they include values between 0 and 1 , while other variables are just exclusively 0 or 1.

Q7

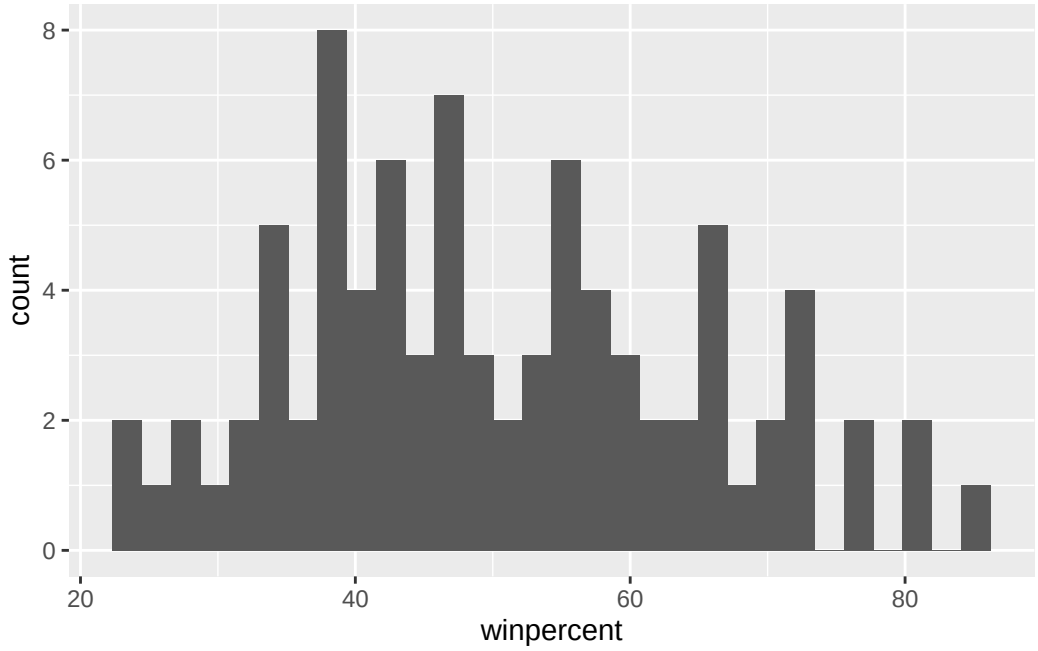
In candy$chocolate column, 0 means the candy is not chocolaty, 1 means the candy is chocolaty.

```
library(ggplot2)
```

Q8

```
ggplot(candy, aes(winpercent))+
  geom_histogram()
```

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



Q9

The distribution is not symmetrical. Is is positively skewed.

Q10

The center of the distribution is below 50%.

Q11

```
choc <- candy$winpercent[as.logical(candy$chocolate)]
fruit <- candy$winpercent[as.logical(candy$fruity)]
mean(choc)>mean(fruit)
```

```
[1] TRUE
```

On average chocolate candy is higher ranked than fruit candy.

Q12

```
t.test(choc, fruit)
```

```
    Welch Two Sample t-test

data:  choc 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
```

The p value is smaller than 0.001. The difference is statistically significant.

```
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
```

Q13

```
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
```

Above are the 5 least liked candy type in this set.

Q14

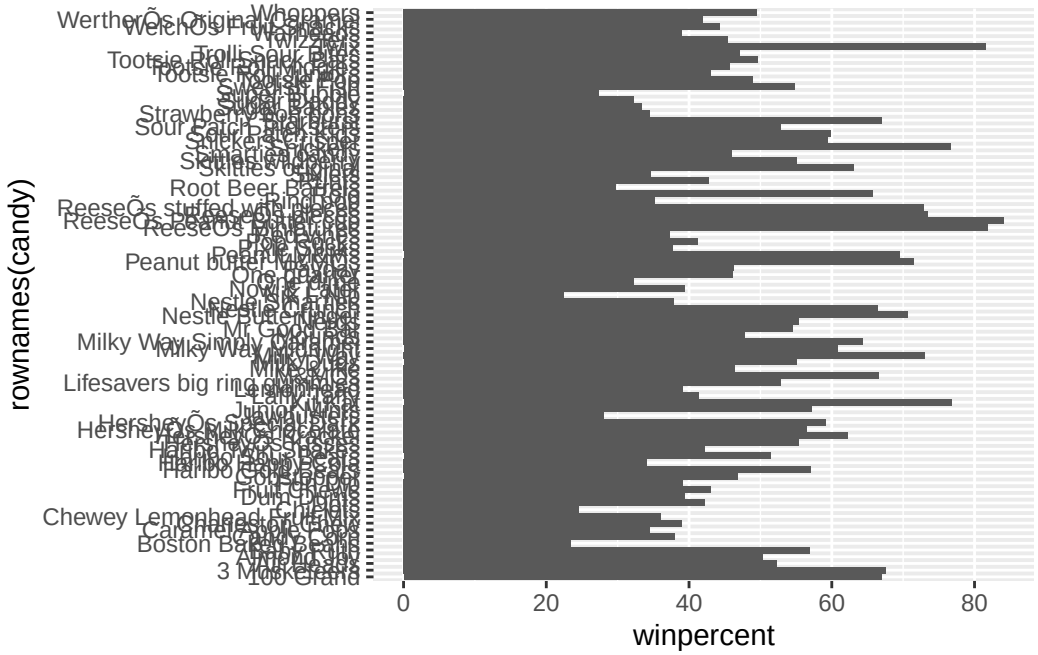
```
candy %>% arrange(winpercent) %>% tail(5)
```

```
                           chocolate fruity caramel peanutyalmondy nougat
Snickers                           1      0       1              1      1
Kit Kat                            1      0       0              0      0
Twix                               1      0       1              0      0
ReeseÕs Miniatures                 1      0       0              1      0
ReeseÕs 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
ReeseÕs Miniatures                         0    0   0        0        0.034
ReeseÕs 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
ReeseÕs Miniatures                0.279   81.86626
ReeseÕs Peanut Butter cup         0.651   84.18029
```

Above are the 5 favorite candy types in this set.

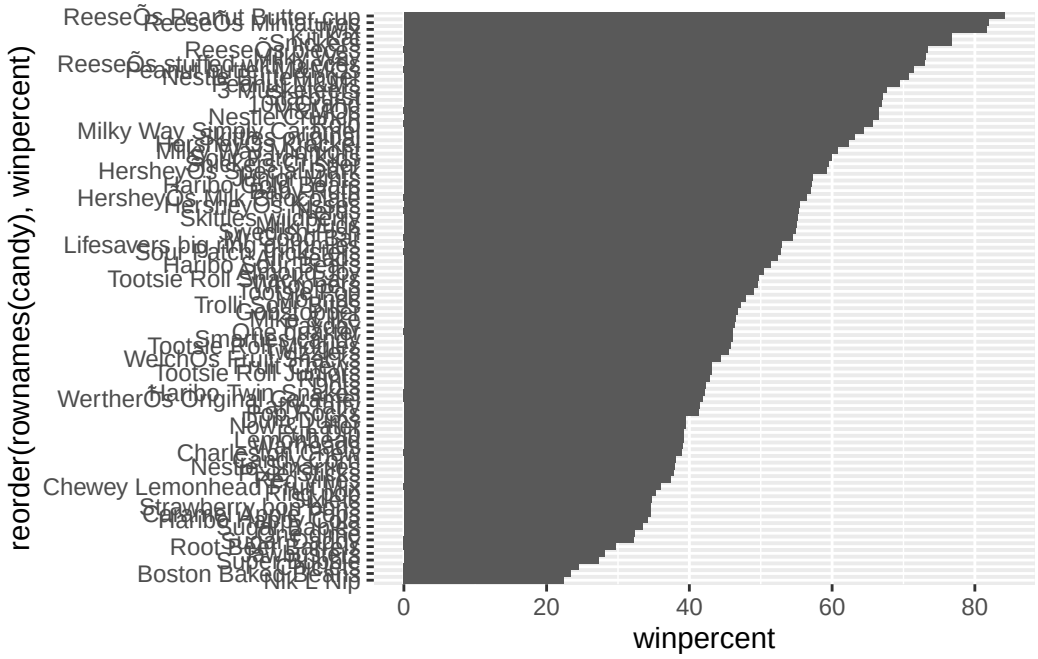
Q15

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



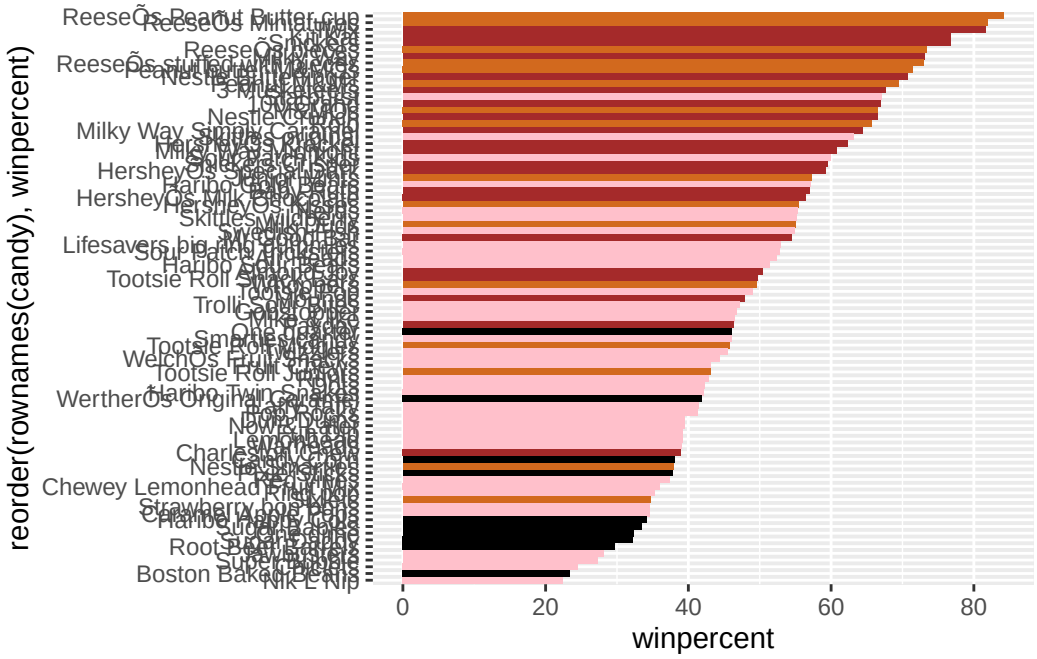
Q16

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




```
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

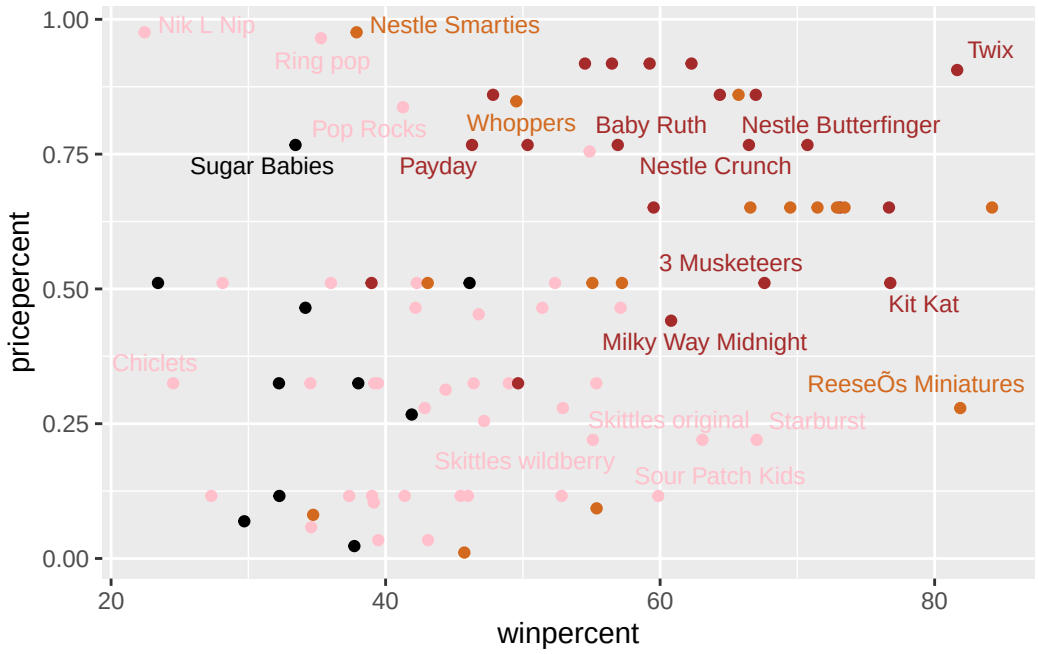
The lowest ranked chocolate candy is sixlets.

Q18

The highest ranked fruit candy is Starburst.

```
library(ggrepel)

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

```
ord <- order(candy$winpercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

```
                          pricepercent winpercent
ReeseÕs Peanut Butter cup        0.651   84.18029
ReeseÕs Miniatures               0.279   81.86626
Twix                             0.906   81.64291
Kit Kat                          0.511   76.76860
Snickers                         0.651   76.67378
```

ReeseÕs Miniatures is the highest ranked in terms of winpercent for the least money

Q20

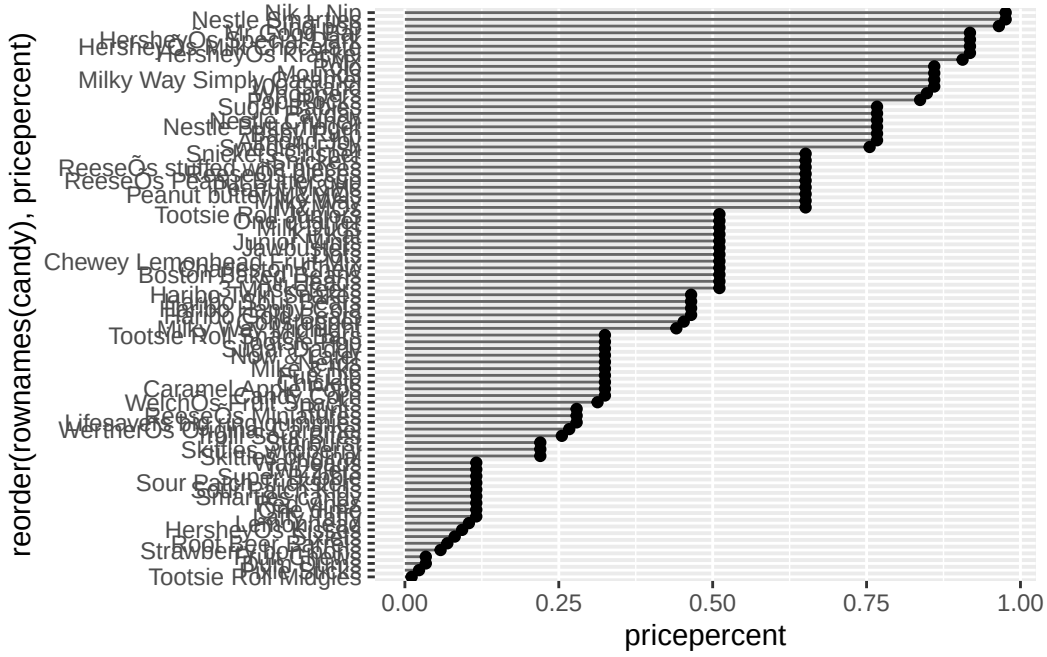
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

```
                        pricepercent winpercent
Nik L Nip                      0.976   22.44534
Nestle Smarties                0.976   37.88719
Ring pop                       0.965   35.29076
HersheyÕs Krackel              0.918   62.28448
HersheyÕs Milk Chocolate       0.918   56.49050
```

Above are the 5 most expensive candy type and Nik L Nip is the least popular one.

Q21

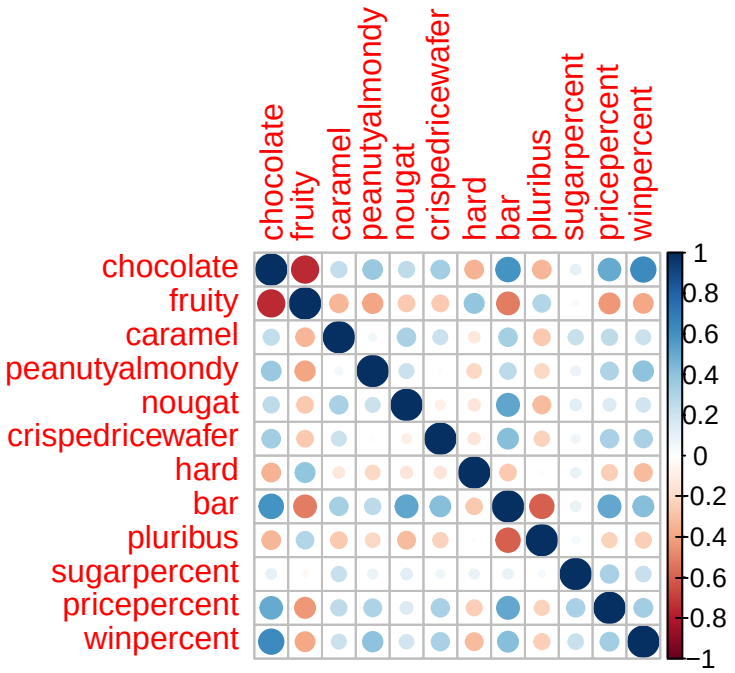
```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```




```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



Q22

Chocolate and fruity are the most anti-correlated, the second one is pluribus and bar.

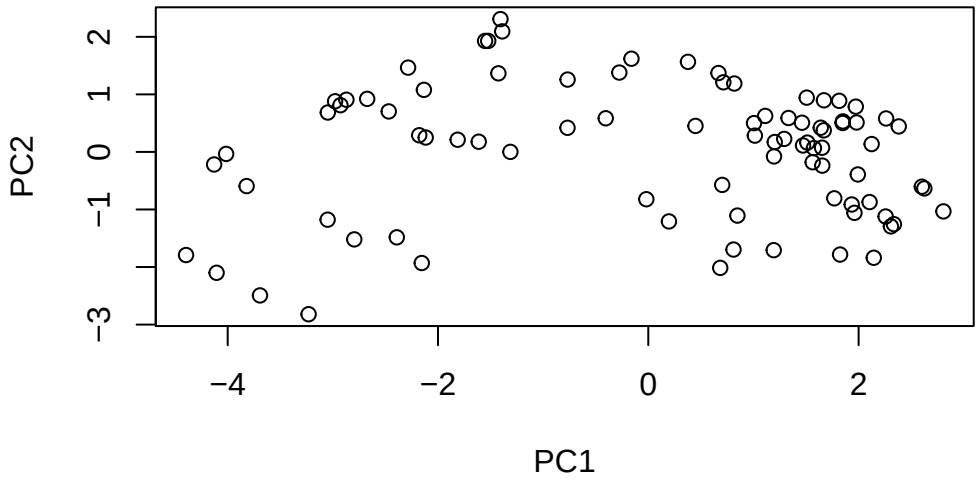
Q23

Chocolate and winpercent are the most positively correlated

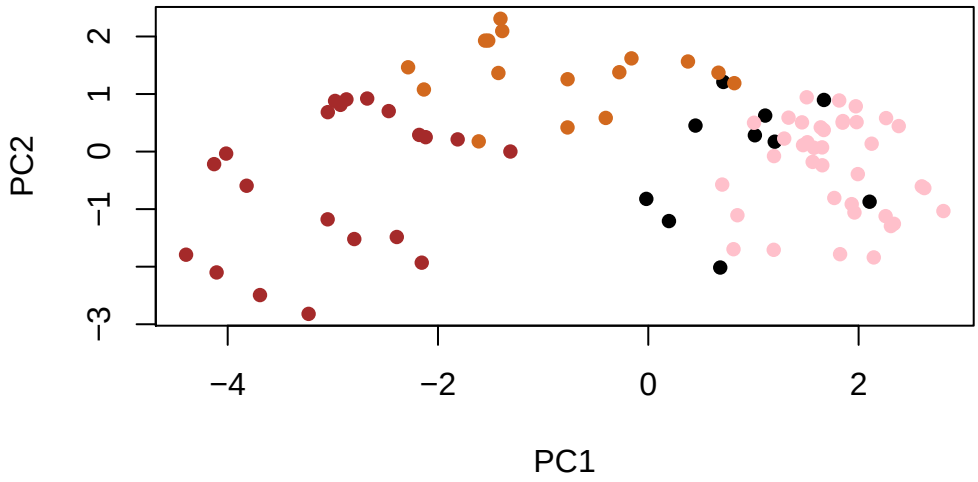
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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
```

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



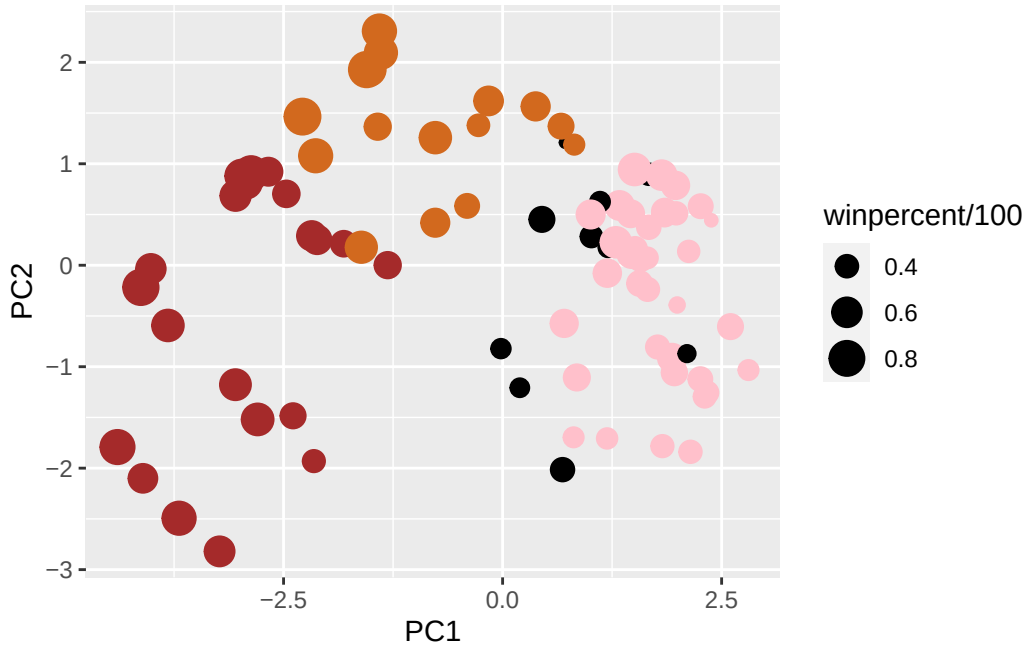

```
plot(pca$x[,1:2], col=my_cols, pch=16)
```




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

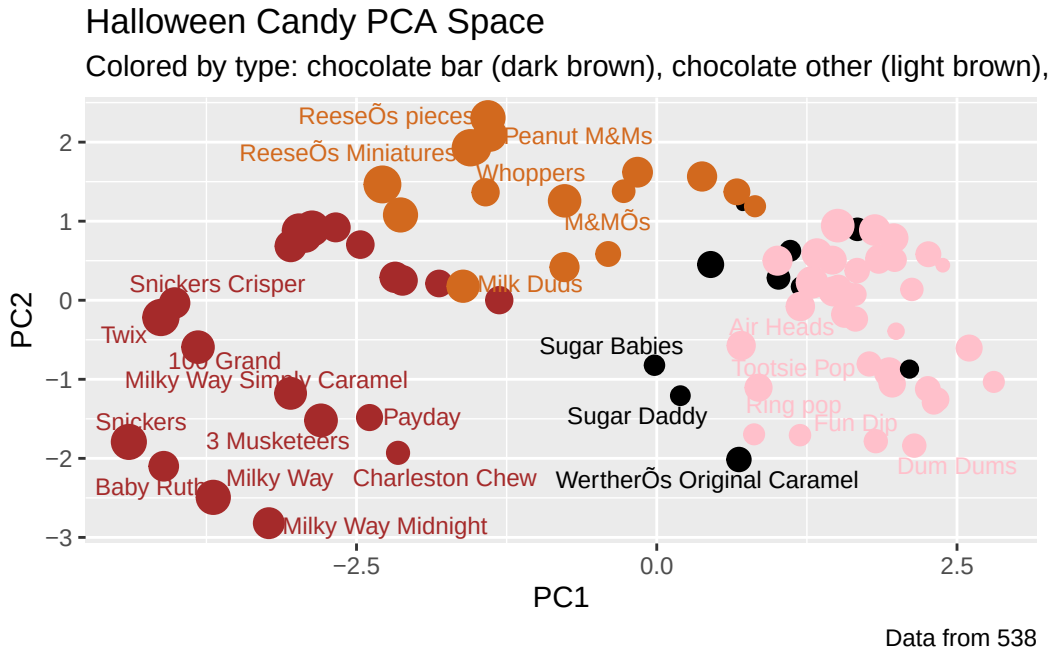
```
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
```

```
        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: 60 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

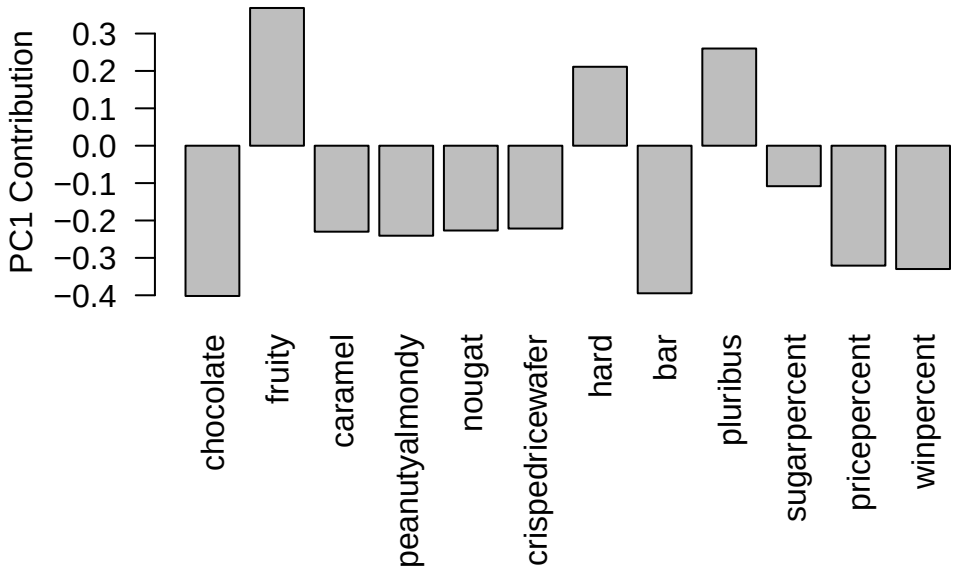



```
library(plotly)
```

```
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)
```

```
par(mar=c(8,4,2,2))
barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```



Q24.

Fruity, pluribus are picked up by PC1 most strongly in the positive direction. This make sense because fruity candy usually come in a bag with multiple candies. Hard is also picked up in the positive direction, but not that strongly.