class10: halloween candy mini project

Sneha Patel

Background

In this mini-project we will examine 538 Halloween Candy data.

First step is read the data

```
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
```

```
nrow(candy)
```

```
[1] 85
```

Q1.

85 types of candy in the dataset

```
table(candy$fruity)
```

```
 0  1
47 38
```

Q2

There are 38 fruity candy types

```
rownames(candy)
```

```
 [1] "100 Grand"                 "3 Musketeers"
 [3] "One dime"                  "One quarter"
 [5] "Air Heads"                 "Almond Joy"
 [7] "Baby Ruth"                 "Boston Baked Beans"
 [9] "Candy Corn"                "Caramel Apple Pops"
[11] "Charleston Chew"           "Chewey Lemonhead Fruit Mix"
[13] "Chiclets"                  "Dots"
[15] "Dum Dums"                  "Fruit Chews"
[17] "Fun Dip"                   "Gobstopper"
[19] "Haribo Gold Bears"         "Haribo Happy Cola"
[21] "Haribo Sour Bears"         "Haribo Twin Snakes"
[23] "HersheyÕs Kisses"          "HersheyÕs Krackel"
[25] "HersheyÕs Milk Chocolate"  "HersheyÕs Special Dark"
[27] "Jawbusters"                "Junior Mints"
[29] "Kit Kat"                   "Laffy Taffy"
[31] "Lemonhead"                 "Lifesavers big ring gummies"
[33] "Peanut butter M&MÕs"       "M&MÕs"
[35] "Mike & Ike"                "Milk Duds"
[37] "Milky Way"                 "Milky Way Midnight"
[39] "Milky Way Simply Caramel"  "Mounds"
[41] "Mr Good Bar"               "Nerds"
[43] "Nestle Butterfinger"       "Nestle Crunch"
[45] "Nik L Nip"                 "Now & Later"
```

```
[47] "Payday"                       "Peanut M&Ms"
[49] "Pixie Sticks"                 "Pop Rocks"
[51] "Red vines"                    "ReeseÕs Miniatures"
[53] "ReeseÕs Peanut Butter cup"    "ReeseÕs pieces"
[55] "ReeseÕs stuffed with pieces" "Ring pop"
[57] "Rolo"                         "Root Beer Barrels"
[59] "Runts"                        "Sixlets"
[61] "Skittles original"            "Skittles wildberry"
[63] "Nestle Smarties"              "Smarties candy"
[65] "Snickers"                     "Snickers Crisper"
[67] "Sour Patch Kids"              "Sour Patch Tricksters"
[69] "Starburst"                    "Strawberry bon bons"
[71] "Sugar Babies"                 "Sugar Daddy"
[73] "Super Bubble"                 "Swedish Fish"
[75] "Tootsie Pop"                  "Tootsie Roll Juniors"
[77] "Tootsie Roll Midgies"         "Tootsie Roll Snack Bars"
[79] "Trolli Sour Bites"            "Twix"
[81] "Twizzlers"                    "Warheads"
[83] "WelchÕs Fruit Snacks"         "WertherÕs Original Caramel"
[85] "Whoppers"
```

```
candy["HersheyÕs Special Dark", ]$winpercent
```

```
[1] 59.23612
```

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

```
[1] 76.7686
```

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

```
[1] 49.6535
```

Q3

My favorite candy is Hershey's Special Dark and the win percent is 59.2%

Q4

Kit Kat's win percent is 76.77%

Q5

Tootsie Roll Snack Bars win percent is 49.65%

```
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

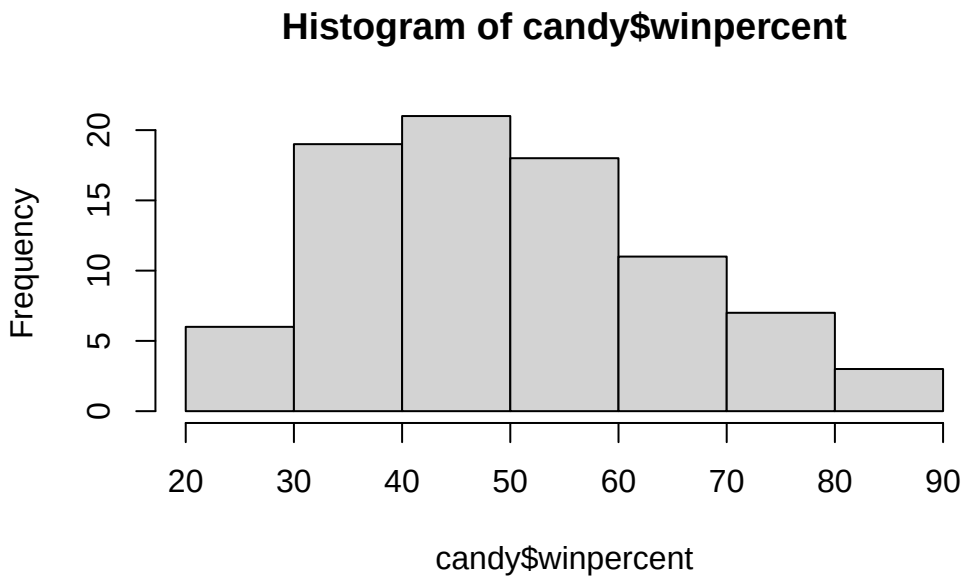
winpercent seems to be on a different scale

Q7

The 1s and 0s in the chocolate column represent true or false aka if it is chocolate or not

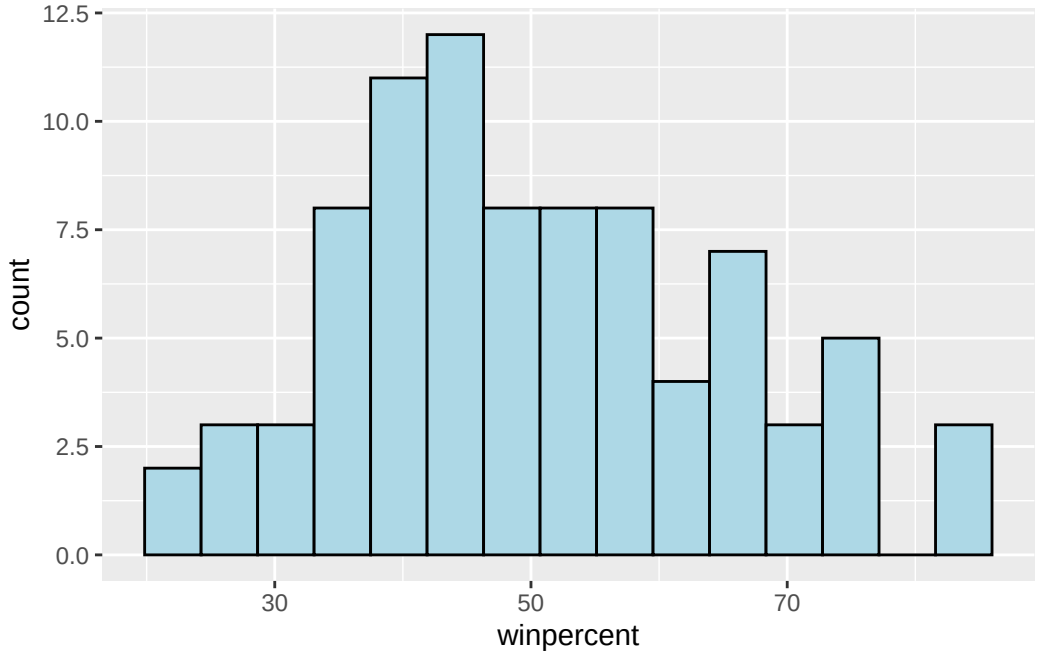
Q8

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




```
library(ggplot2)

ggplot(candy) + aes (x = winpercent) + geom_histogram(bins=15, col="black", fill="lightblu
```



Q9

The distribution of `winpercent` is not symmetrical

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

```
[1] 47.82975
```

Q10

The center of the distribution is below 50%

```
#choc_avg <- mean(candy$winpercent(chocolate==1))
choc_true <-  as.logical(candy$chocolate)
choc_all <- candy[choc_true,]$winpercent
choc_avg <- mean(choc_all)

fruity_true <-  as.logical(candy$fruity)
fruity_all <-candy[fruity_true,]$winpercent
```

```r
fruity_avg <- mean(fruity_all)

choc_avg > fruity_avg
```

```
[1] TRUE
```

Q11

Chocolate candies are ranked higher than fruit candy

```r
t.test(choc_all, fruity_all)
```

```

    Welch Two Sample t-test

data:  choc_all and fruity_all
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
```

Q12

The difference is not statistically significant

```r
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
```

Q13

The 5 least liked candies are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters

```
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
```

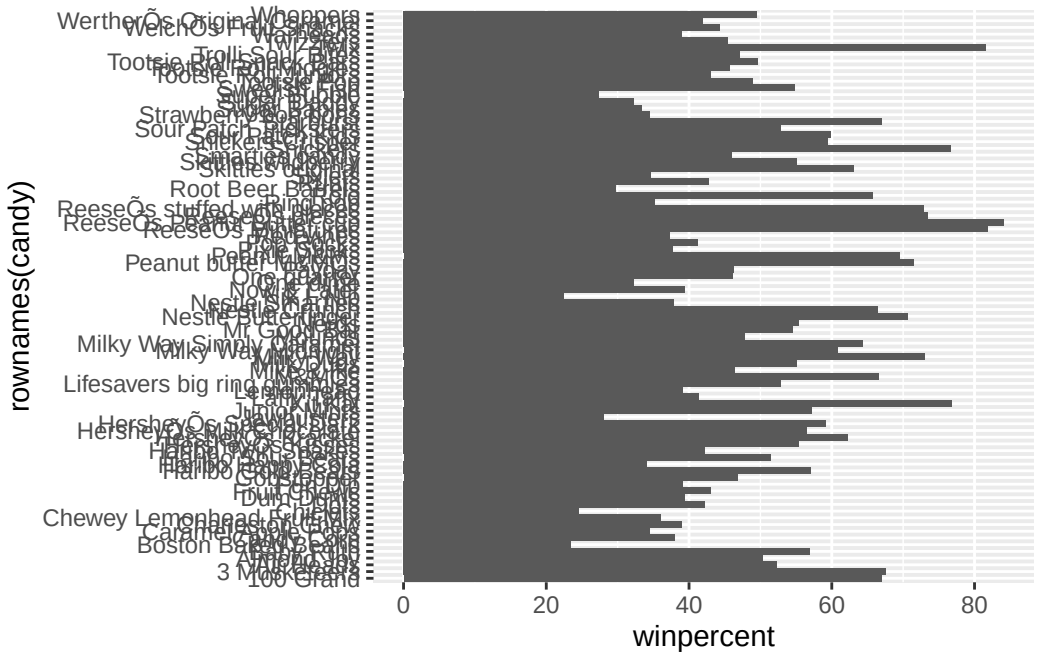
```
Snickers                            0    0    1        0        0.546
Kit Kat                             1    0    1        0        0.313
Twix                                1    0    1        0        0.546
ReeseÕs Miniatures                  0    0    0        0        0.034
ReeseÕs Peanut Butter 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
```

Q14

The top 5 candies are Snickers, Kit Kat, Twix, Reese's Miniatures, and Reese's Peanut Butter cup

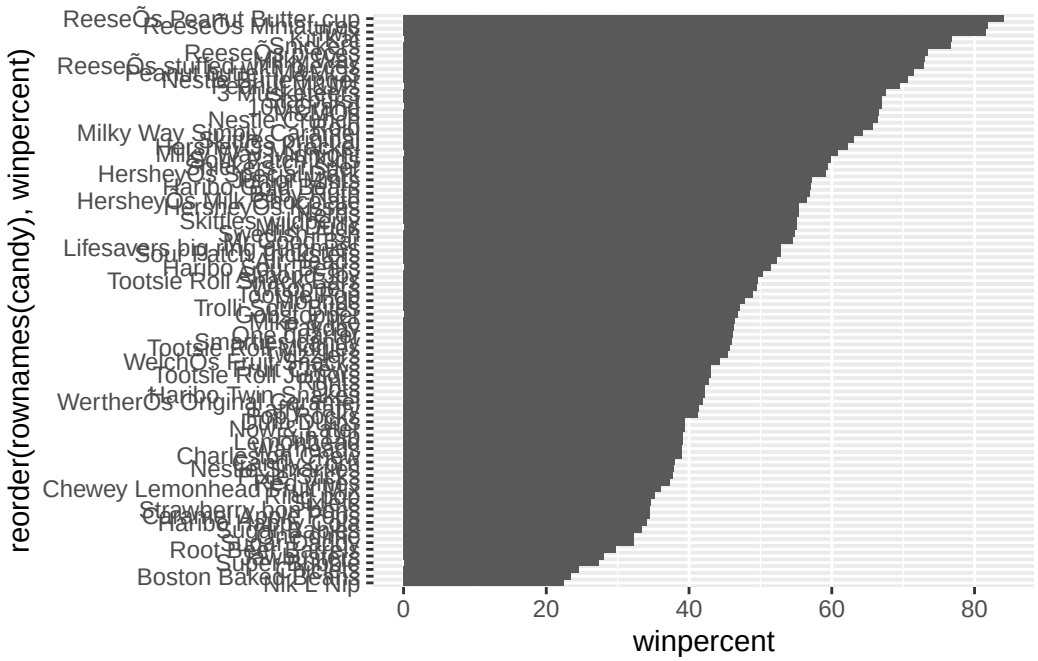
```
library(ggplot2)

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



```
library(ggplot2)

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




```
ggsave("mybarplot.png")
```

```
Saving 5.5 x 3.5 in image
```

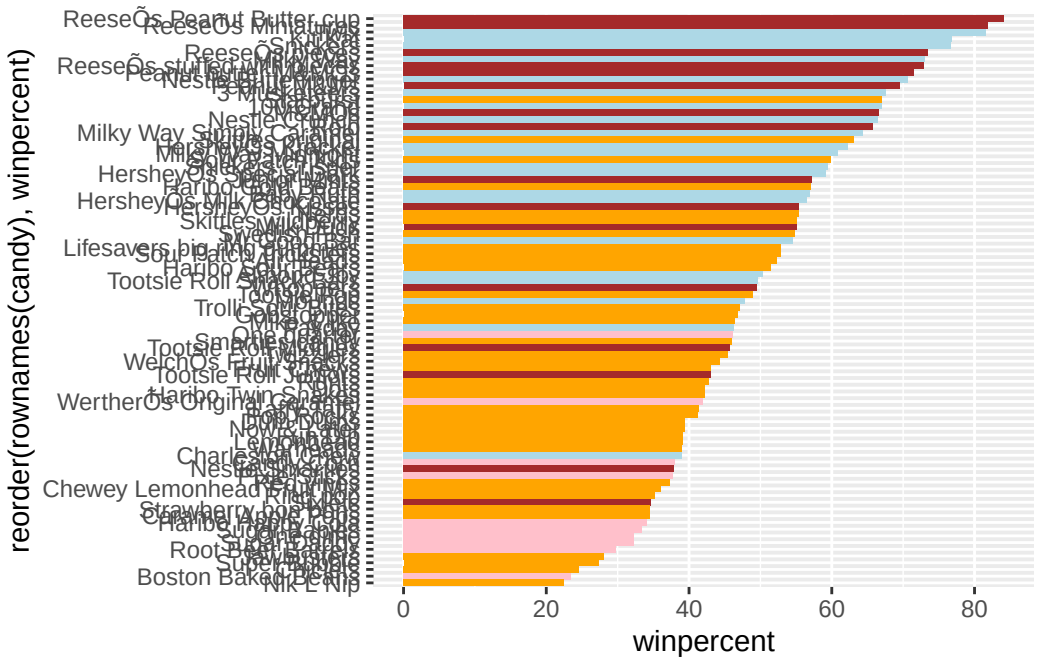
Adding color

Let's setup a color vector (that signifies candy type) that we can then use for some future plots. We start by making a vector of all black values (one for each candy). Then we overwrite chocolate (for chocolate candy), brown (for candy bars) and red (for fruity candy) values.

```
my_cols <- rep("pink", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "brown"
my_cols[as.logical(candy$bar)] = "lightblue"
my_cols[as.logical(candy$fruity)] = "orange"
```

```
library(ggplot2)

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



Q17

Worst ranked chocolate is sixlets

Q18

Best ranked fruity candy is starburst

Taking a look at pricepercent

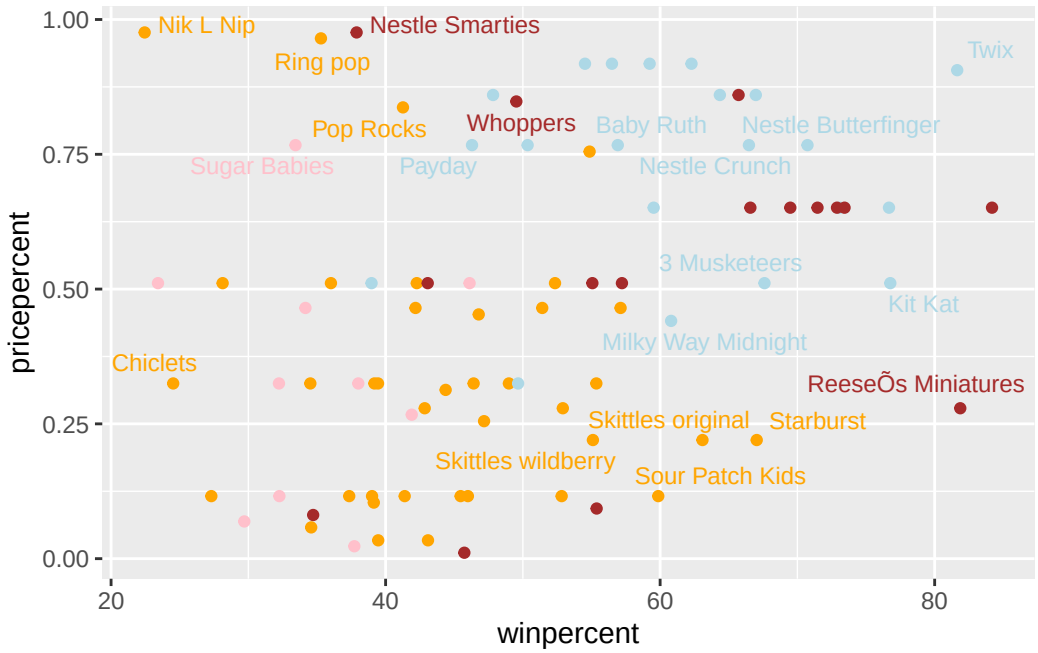
What about value for money? What is the the best candy for the least money? One way to get at this would be to make a plot of winpercent vs the pricepercent variable. The pricepercent variable records the percentile rank of the candy's price against all the other candies in the dataset. Lower vales are less expensive and high values more expensive.

To this plot we will add text labels so we can more easily identify a given candy. There is a regular geom_label() that comes with ggplot2. However, as there are quite a few candys in our dataset lots of these labels will be overlapping and hard to read. To help with this we can use the geom_text_repel() function from the ggrepel package.

```
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
```




```
#highest winpercent, lowest pricepercent
ord <- order(candy$pricepercent, decreasing = FALSE)
head(candy[ord,c(11,12)], n=5 )
```

```
                pricepercent winpercent
```

```
Tootsie Roll Midgies        0.011   45.73675
Pixie Sticks                0.023   37.72234
Dum Dums                    0.034   39.46056
Fruit Chews                 0.034   43.08892
Strawberry bon bons         0.058   34.57899
```

Q19

Tootsie roll midgies is the biggest bang for your buck

```
#lowest winpercent, highest pricepercent
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
```

Q20

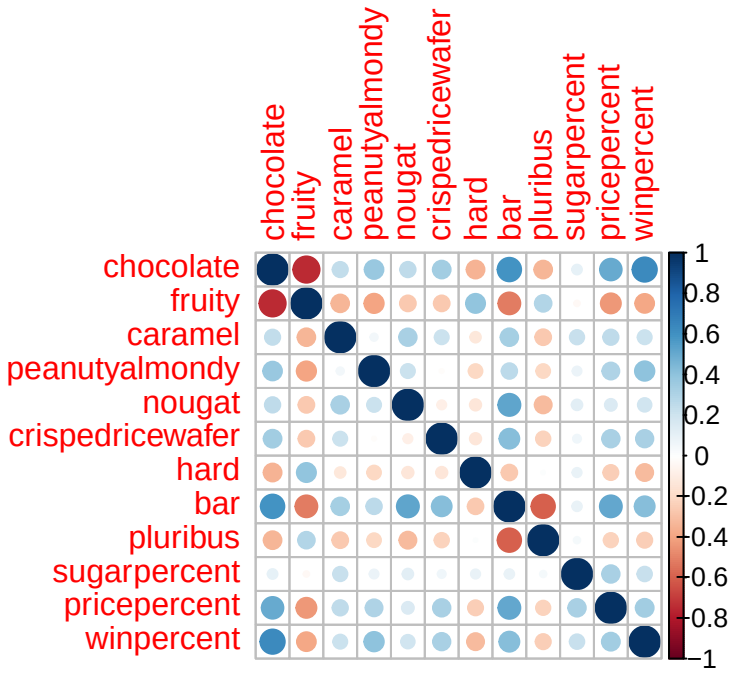
Worst expensive candy is Nik L Nip

Exploring the correlation structure

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



Q22

Fruit and Chocolate are anti-correlated

Q23

Chocolate and bar/winpercent are the most positively correlated

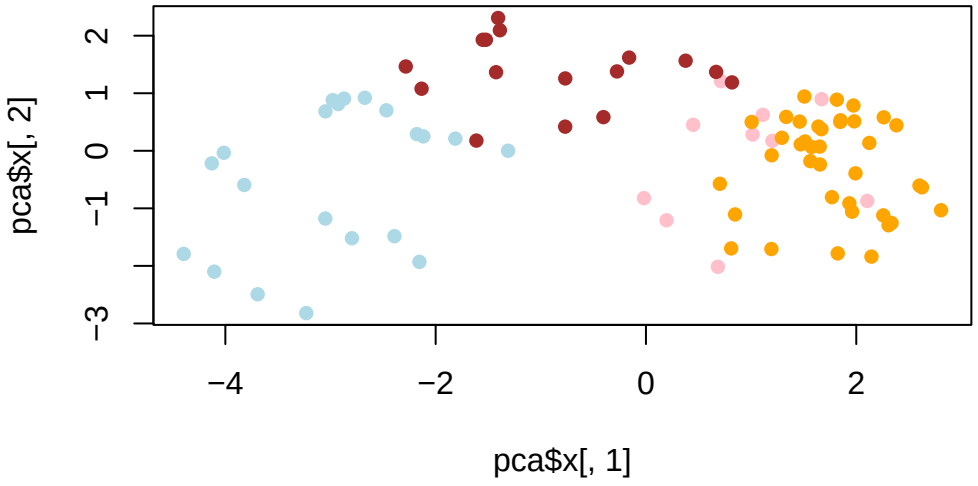
PCA

```
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
```

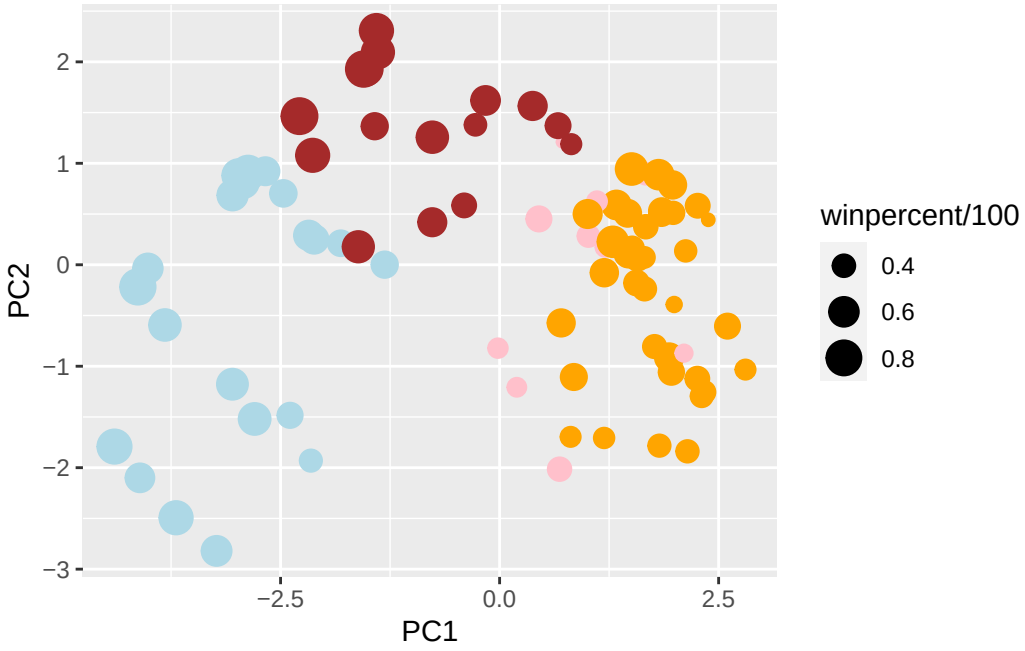
```
                            PC8     PC9    PC10    PC11    PC12
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

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



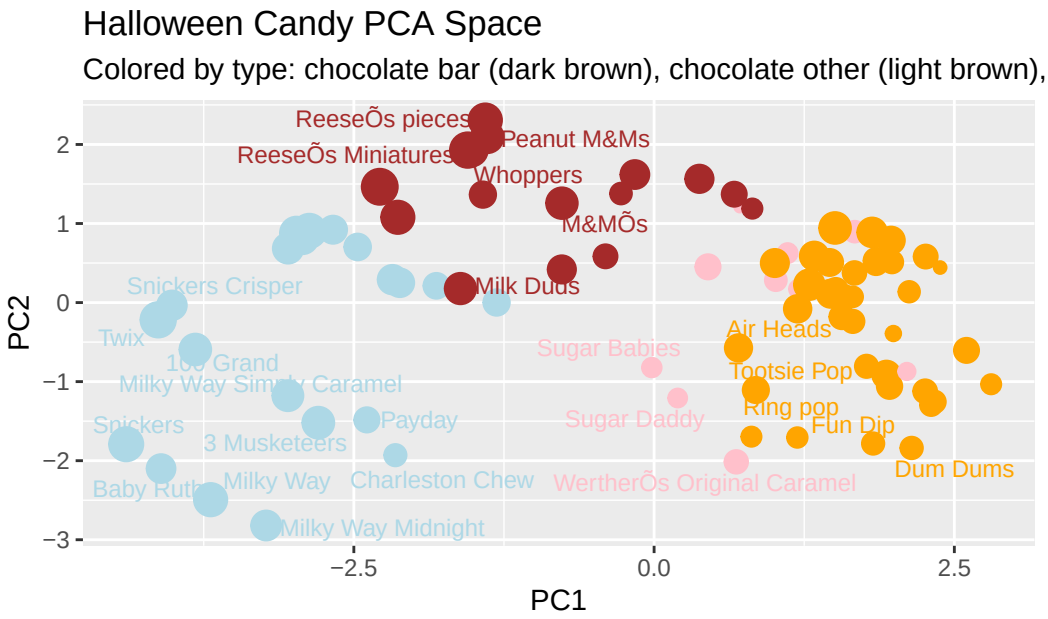

```r
my_data <- cbind(candy, pca$x[,1:3])

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
```



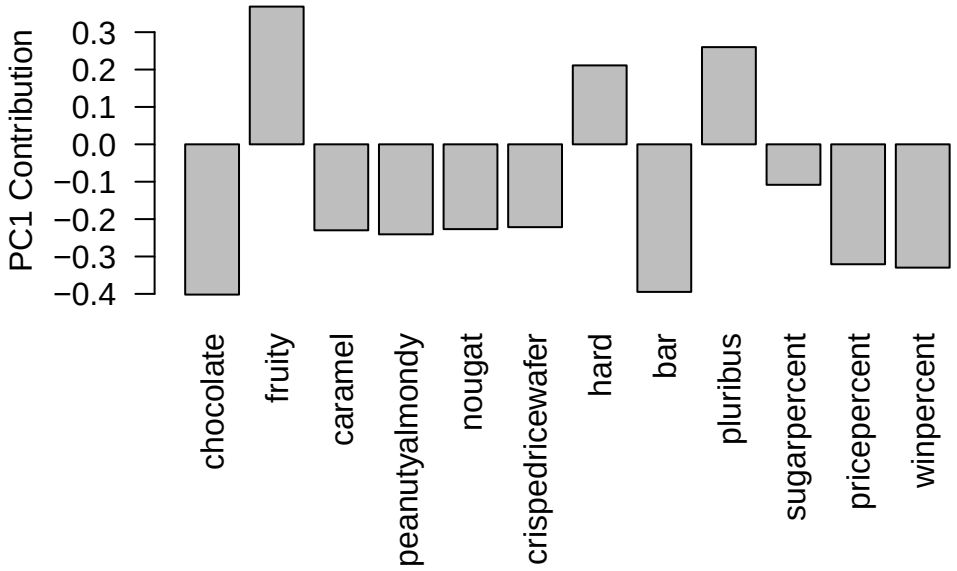

```
p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
       caption="Data from 538")
```

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



Data from 538

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



Q24

Fruity, hard, and pluribus have strong PC1 in the positive direction. This makes sense since chocolate isn't hard, and hard candies typically come in individually wrapped bite size pieces