Class10

Gregory Jordan

Table of contents

Background

In this mini project we wille xamine 538 Halloween candy data. what is your favorite candy? what is nougat anyway? and how do you say it in america?

1. Importing the dataset

First step is to read the data:

```
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
```

win percent means this candy was their favorite when asked people

Q1. How many different candy types are in this dataset?

```
cat(nrow(candy),"different types of candy")
```

```
85 different types of candy
```

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

```
cat(sum(candy$fruity),"fruity candy types in the dataset")
```

```
38 fruity candy types in the dataset
```

2. What is your favorite candy?

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

```
#let's see all the different types of candy to find my favorite.
row.names(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"
```

```
#my favorite candy is Twix
candy["Twix",]$winpercent
```

```
[1] 81.64291
```

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
```

```
#install.packages("skimr")
library(skimr)
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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?

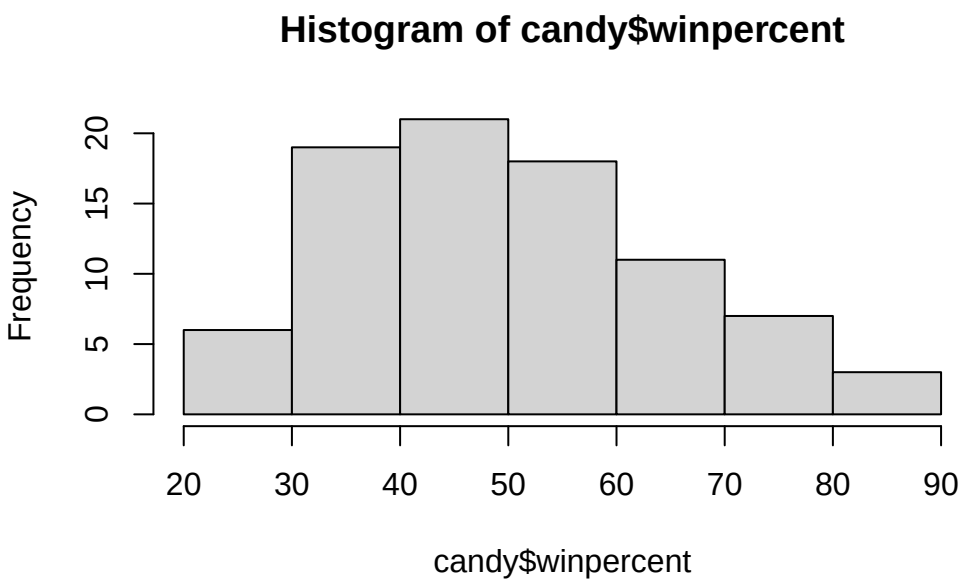
winpercent is on a 1-100% scale while the other variables are on a 0-1 scale.

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

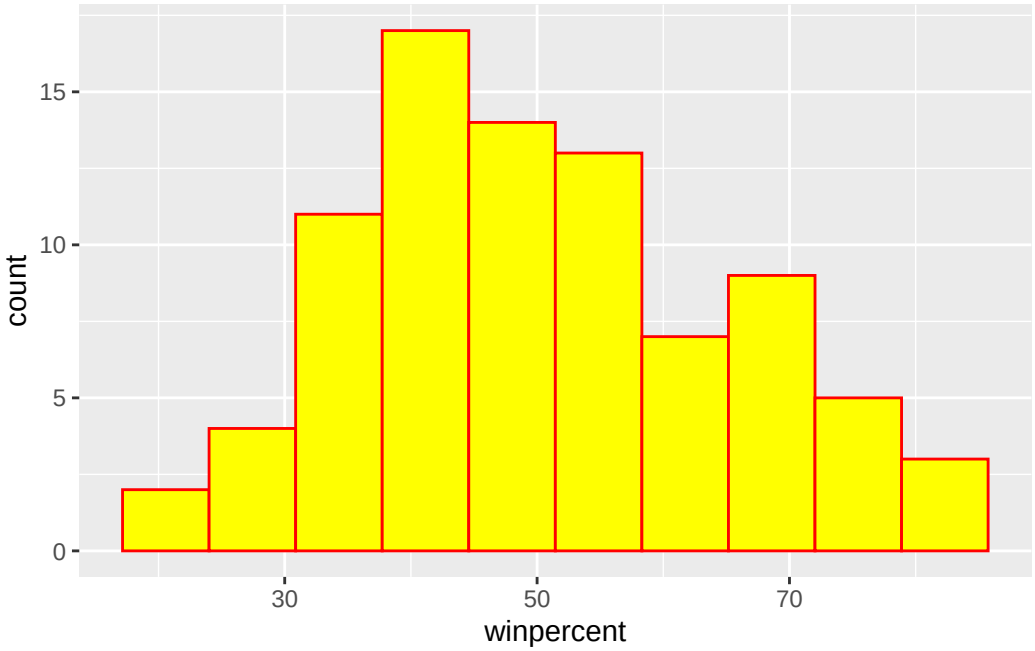
A zero and one represent whether the candey was a chocolate type of candy or not

Q8. Plot a histogram of winpercent values

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




```
#we can use ggplot also to plot the hist and make it fancier
library(ggplot2)

ggplot(candy) + aes(winpercent) + geom_histogram(bins=10,col="red",fill="yellow")
```



Q9. Is the distribution of winpercent values symmetrical?

No. skewed towards <50%

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

Center of the distribution is skewed below 50%

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

```
chocolate.win<- mean(candy$winpercent[as.logical(candy$chocolate)])
cat(chocolate.win, "= mean win % chocolate\n")
```

```
60.92153 = mean win % chocolate
```

```
fruity.win<- mean(candy$winpercent[as.logical(candy$fruity)])
cat(fruity.win,"= mean win $ fruity\n")
```

```
44.11974 = mean win $ fruity
```

```
cat("Chocolate > Fruity?", chocolate.win>fruity.win)
```

```
Chocolate > Fruity? TRUE
```

Q12. Is this difference statistically significant?

```
#student t test to test for significance
t.test(candy$winpercent[as.logical(candy$chocolate)],candy$winpercent[as.logical(candy$fru
```

```
    Welch Two Sample t-test

data:  candy$winpercent[as.logical(candy$chocolate)] and candy$winpercent[as.logical(candy$fi
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 it is stat significant b/c super low p value

3. Overall Candy Rankings

the base R `sort()` and `order()` functions are very useful! dplyr works well too!

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

```
#I like tidyverse/dplyr
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
```

```r
candy.least.liked <- candy %>% arrange(winpercent)
head(candy.least.liked,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
```

```r
#can also use base R instead of tidyverse
inds <- order(candy$winpercent)
head(candy[inds,],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
```

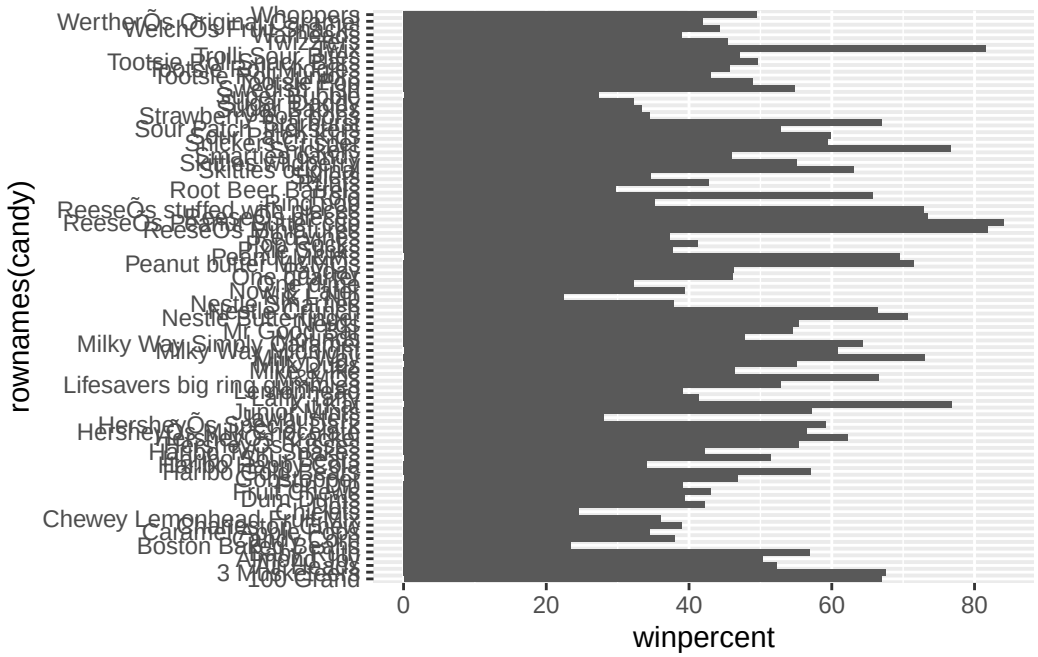
Q14. What are the top 5 all time favorite candy types out of this set?

```
#using base R
inds <- order(candy$winpercent)
tail(candy[inds,],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
```

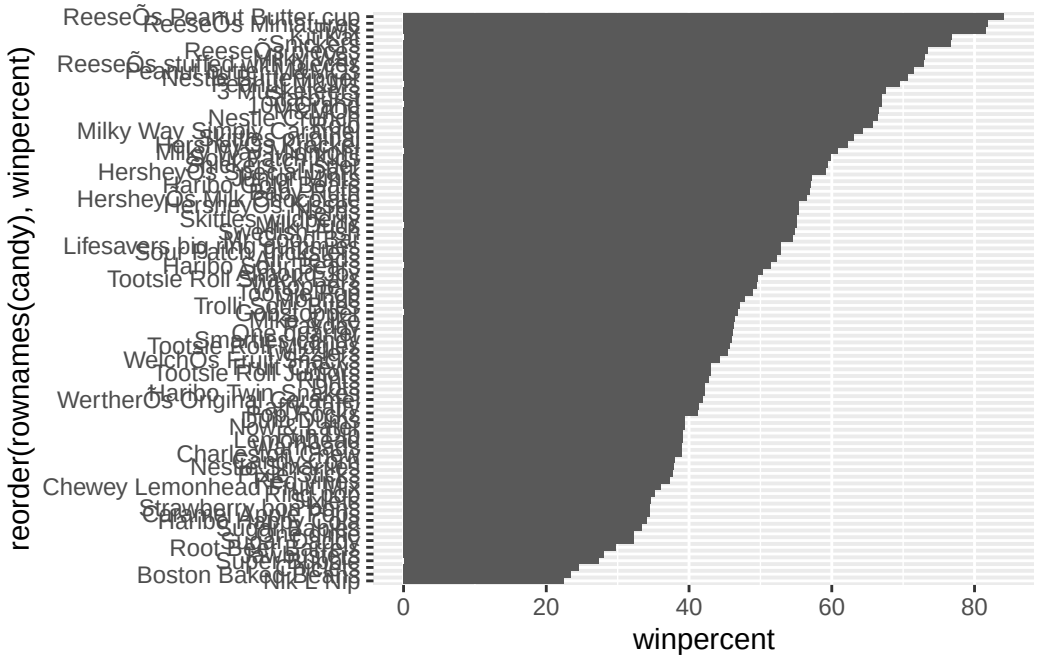
Q15. Make a first barplot of candy ranking based on winpercent values.

```
#library(ggplot2)
ggplot(data=candy) + aes(winpercent,rownames(candy)) + geom_col()
```



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

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



you can use `ggsave()` to save/edit dimensions and save your most recent plot if you want

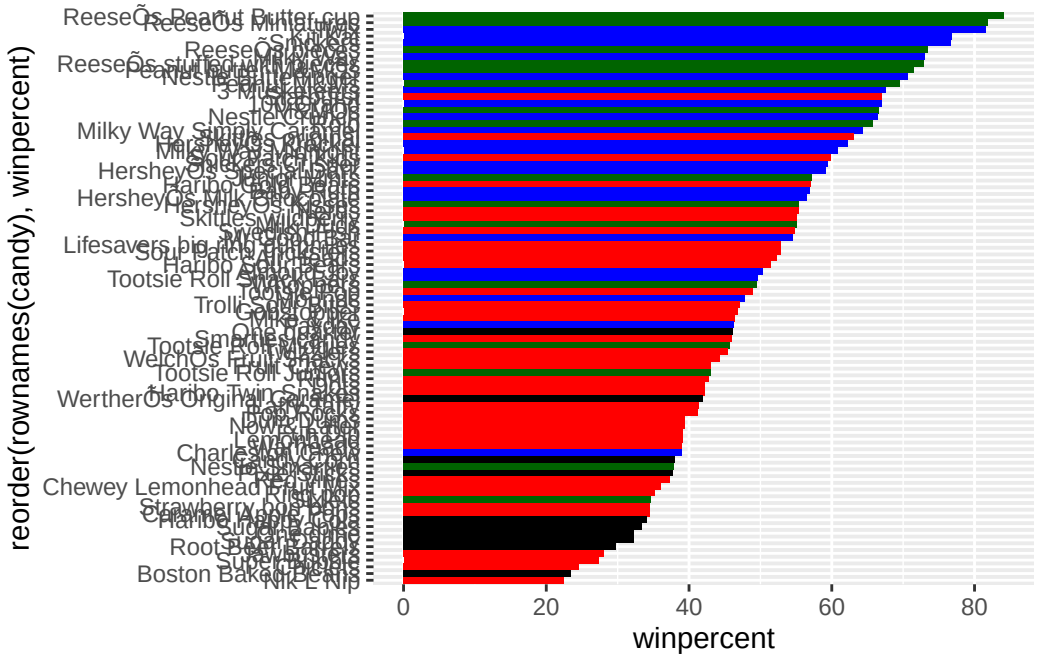
let's add color. use a color vector that we can then use to color candies by descriptions (like chocolate and stuff)

```
#start by making a vector of all black color as long as the different types of candies
my_cols <- rep("black",nrow(candy))
#my_cols
#then overwrite the vector to rename colors based off candy
my_cols[as.logical(candy$chocolate)] <- "darkgreen"
my_cols[as.logical(candy$bar)] <- "blue"
my_cols[as.logical(candy$fruity)] <- "red"
my_cols
```

```
 [1] "blue"      "blue"      "black"     "black"     "red"       "blue"
 [7] "blue"      "black"     "black"     "red"       "blue"      "red"
[13] "red"       "red"       "red"       "red"       "red"       "red"
[19] "red"       "black"     "red"       "red"       "darkgreen" "blue"
[25] "blue"      "blue"      "red"       "darkgreen" "blue"      "red"
[31] "red"       "red"       "darkgreen" "darkgreen" "red"       "darkgreen"
[37] "blue"      "blue"      "blue"      "blue"      "blue"      "red"
[43] "blue"      "blue"      "red"       "red"       "blue"      "darkgreen"
```

```
[49] "black"     "red"       "red"       "darkgreen" "darkgreen" "darkgreen"
[55] "darkgreen" "red"       "darkgreen" "black"     "red"       "darkgreen"
[61] "red"       "red"       "darkgreen" "red"       "blue"      "blue"
[67] "red"       "red"       "red"       "red"       "black"     "black"
[73] "red"       "red"       "red"       "darkgreen" "darkgreen" "blue"
[79] "red"       "blue"      "red"       "red"       "red"       "black"
[85] "darkgreen"
```

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



Q17. What is the worst ranked chocolate candy?

sixlets

```
candy.worst.chocolate <- candy %>% filter(chocolate==1) %>% arrange(winpercent)
candy.worst.chocolate[1,]
```

```
        chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
Sixlets         1      0       0              0      0                 0    0
        bar pluribus sugarpercent pricepercent winpercent
Sixlets   0        1         0.22        0.081     34.722
```

Q18. What is the best ranked fruity candy?

starburst

```
candy.best.fruity <- candy %>% filter(fruity==1) %>% arrange(desc(winpercent))
candy.best.fruity[1,]
```

```
          chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
Starburst         0      1       0              0      0                0    0
          bar pluribus sugarpercent pricepercent winpercent
Starburst   0        1        0.151         0.22   67.03763
```

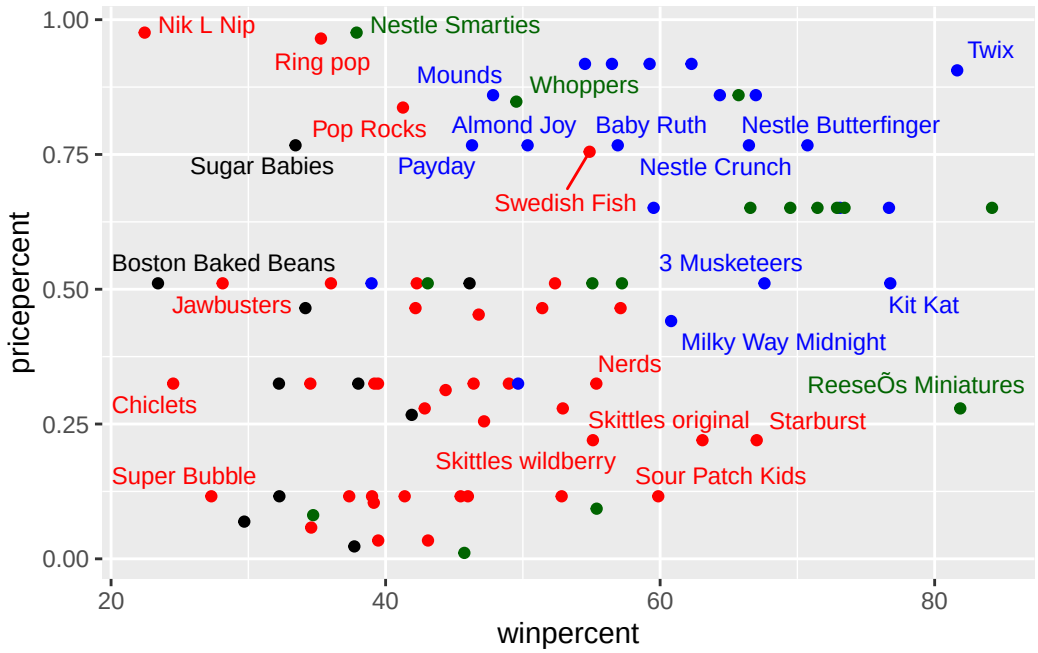
4. Taking a look at pricepercent

what about value for money? what is the besat candy for the least money?

one way to get at this would be to make a plot of `winpercent` vs the `pricepercent` variables

```
#install.packages(ggrepel)
#ggrepel to make labels not overlap
library(ggrepel)
ggplot(candy) + aes(winpercent,pricepercent,label=rownames(candy)) + geom_point(col=my_col
```

```
Warning: ggrepel: 58 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?

```
candy.bangforbuck <- candy %>% mutate(bangforbuck=winpercent/pricepercent) %>% arrange(des
candy.bangforbuck
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Tootsie Roll Midgies	1	0	0	0	0
Pixie Sticks	0	0	0	0	0
Fruit Chews	0	1	0	0	0
Dum Dums	0	1	0	0	0
Strawberry bon bons	0	1	0	0	0
HersheyÕs Kisses	1	0	0	0	0
Sour Patch Kids	0	1	0	0	0
Sour Patch Tricksters	0	1	0	0	0
Root Beer Barrels	0	0	0	0	0
Sixlets	1	0	0	0	0
Smarties candy	0	1	0	0	0
Twizzlers	0	1	0	0	0
Lemonhead	0	1	0	0	0
Laffy Taffy	0	1	0	0	0
Warheads	0	1	0	0	0

Red vines	0	1	0	0	0
Starburst	0	1	0	0	0
ReeseÕs Miniatures	1	0	0	1	0
Skittles original	0	1	0	0	0
One dime	0	0	0	0	0
Skittles wildberry	0	1	0	0	0
Super Bubble	0	1	0	0	0
Lifesavers big ring gummies	0	1	0	0	0
Trolli Sour Bites	0	1	0	0	0
Nerds	0	1	0	0	0
WertherÕs Original Caramel	0	0	1	0	0
Runts	0	1	0	0	0
Tootsie Roll Snack Bars	1	0	0	0	0
Tootsie Pop	1	1	0	0	0
Kit Kat	1	0	0	0	0
Mike & Ike	0	1	0	0	0
WelchÕs Fruit Snacks	0	1	0	0	0
Milky Way Midnight	1	0	1	0	1
3 Musketeers	1	0	0	0	1
ReeseÕs Peanut Butter cup	1	0	0	1	0
Haribo Gold Bears	0	1	0	0	0
Now & Later	0	1	0	0	0
Fun Dip	0	1	0	0	0
Snickers	1	0	1	1	1
Candy Corn	0	0	0	0	0
ReeseÕs pieces	1	0	0	1	0
Milky Way	1	0	1	0	1
Junior Mints	1	0	0	0	0
ReeseÕs stuffed with pieces	1	0	0	1	0
Haribo Sour Bears	0	1	0	0	0
Peanut butter M&MÕs	1	0	0	1	0
Milk Duds	1	0	1	0	0
Peanut M&Ms	1	0	0	1	0
Caramel Apple Pops	0	1	1	0	0
Gobstopper	0	1	0	0	0
Air Heads	0	1	0	0	0
M&MÕs	1	0	0	0	0
Sugar Daddy	0	0	1	0	0
Nestle Butterfinger	1	0	0	1	0
Snickers Crisper	1	0	1	1	0
Haribo Twin Snakes	0	1	0	0	0
One quarter	0	0	0	0	0
Twix	1	0	1	0	0

Nestle Crunch	1	0	0	0	0
Tootsie Roll Juniors	1	0	0	0	0
Dots	0	1	0	0	0
100 Grand	1	0	1	0	0
Rolo	1	0	1	0	0
Charleston Chew	1	0	0	0	1
Chiclets	0	1	0	0	0
Milky Way Simply Caramel	1	0	1	0	0
Baby Ruth	1	0	1	1	1
Haribo Happy Cola	0	0	0	0	0
Swedish Fish	0	1	0	0	0
Chewey Lemonhead Fruit Mix	0	1	0	0	0
HersheyÕs Krackel	1	0	0	0	0
Almond Joy	1	0	0	1	0
HersheyÕs Special Dark	1	0	0	0	0
HersheyÕs Milk Chocolate	1	0	0	0	0
Payday	0	0	0	1	1
Mr Good Bar	1	0	0	1	0
Whoppers	1	0	0	0	0
Mounds	1	0	0	0	0
Jawbusters	0	1	0	0	0
Pop Rocks	0	1	0	0	0
Boston Baked Beans	0	0	0	1	0
Sugar Babies	0	0	1	0	0
Nestle Smarties	1	0	0	0	0
Ring pop	0	1	0	0	0
Nik L Nip	0	1	0	0	0

	crispedricewafer	hard	bar	pluribus	sugarpercent
Tootsie Roll Midgies	0	0	0	1	0.174
Pixie Sticks	0	0	0	1	0.093
Fruit Chews	0	0	0	1	0.127
Dum Dums	0	1	0	0	0.732
Strawberry bon bons	0	1	0	1	0.569
HersheyÕs Kisses	0	0	0	1	0.127
Sour Patch Kids	0	0	0	1	0.069
Sour Patch Tricksters	0	0	0	1	0.069
Root Beer Barrels	0	1	0	1	0.732
Sixlets	0	0	0	1	0.220
Smarties candy	0	1	0	1	0.267
Twizzlers	0	0	0	0	0.220
Lemonhead	0	1	0	0	0.046
Laffy Taffy	0	0	0	0	0.220
Warheads	0	1	0	0	0.093

Red vines	0	0	0	1	0.581
Starburst	0	0	0	1	0.151
ReeseÕs Miniatures	0	0	0	0	0.034
Skittles original	0	0	0	1	0.941
One dime	0	0	0	0	0.011
Skittles wildberry	0	0	0	1	0.941
Super Bubble	0	0	0	0	0.162
Lifesavers big ring gummies	0	0	0	0	0.267
Trolli Sour Bites	0	0	0	1	0.313
Nerds	0	1	0	1	0.848
WertherÕs Original Caramel	0	1	0	0	0.186
Runts	0	1	0	1	0.872
Tootsie Roll Snack Bars	0	0	1	0	0.465
Tootsie Pop	0	1	0	0	0.604
Kit Kat	1	0	1	0	0.313
Mike & Ike	0	0	0	1	0.872
WelchÕs Fruit Snacks	0	0	0	1	0.313
Milky Way Midnight	0	0	1	0	0.732
3 Musketeers	0	0	1	0	0.604
ReeseÕs Peanut Butter cup	0	0	0	0	0.720
Haribo Gold Bears	0	0	0	1	0.465
Now & Later	0	0	0	1	0.220
Fun Dip	0	1	0	0	0.732
Snickers	0	0	1	0	0.546
Candy Corn	0	0	0	1	0.906
ReeseÕs pieces	0	0	0	1	0.406
Milky Way	0	0	1	0	0.604
Junior Mints	0	0	0	1	0.197
ReeseÕs stuffed with pieces	0	0	0	0	0.988
Haribo Sour Bears	0	0	0	1	0.465
Peanut butter M&MÕs	0	0	0	1	0.825
Milk Duds	0	0	0	1	0.302
Peanut M&Ms	0	0	0	1	0.593
Caramel Apple Pops	0	0	0	0	0.604
Gobstopper	0	1	0	1	0.906
Air Heads	0	0	0	0	0.906
M&MÕs	0	0	0	1	0.825
Sugar Daddy	0	0	0	0	0.418
Nestle Butterfinger	0	0	1	0	0.604
Snickers Crisper	1	0	1	0	0.604
Haribo Twin Snakes	0	0	0	1	0.465
One quarter	0	0	0	0	0.011
Twix	1	0	1	0	0.546

Nestle Crunch	1	0	1	0	0.313
Tootsie Roll Juniors	0	0	0	0	0.313
Dots	0	0	0	1	0.732
100 Grand	1	0	1	0	0.732
Rolo	0	0	0	1	0.860
Charleston Chew	0	0	1	0	0.604
Chiclets	0	0	0	1	0.046
Milky Way Simply Caramel	0	0	1	0	0.965
Baby Ruth	0	0	1	0	0.604
Haribo Happy Cola	0	0	0	1	0.465
Swedish Fish	0	0	0	1	0.604
Chewey Lemonhead Fruit Mix	0	0	0	1	0.732
HersheyÕs Krackel	1	0	1	0	0.430
Almond Joy	0	0	1	0	0.465
HersheyÕs Special Dark	0	0	1	0	0.430
HersheyÕs Milk Chocolate	0	0	1	0	0.430
Payday	0	0	1	0	0.465
Mr Good Bar	0	0	1	0	0.313
Whoppers	1	0	0	1	0.872
Mounds	0	0	1	0	0.313
Jawbusters	0	1	0	1	0.093
Pop Rocks	0	1	0	1	0.604
Boston Baked Beans	0	0	0	1	0.313
Sugar Babies	0	0	0	1	0.965
Nestle Smarties	0	0	0	1	0.267
Ring pop	0	1	0	0	0.732
Nik L Nip	0	0	0	1	0.197

	pricepercent	winpercent	bangforbuck
Tootsie Roll Midgies	0.011	45.73675	4157.88618
Pixie Sticks	0.023	37.72234	1640.10157
Fruit Chews	0.034	43.08892	1267.32122
Dum Dums	0.034	39.46056	1160.60452
Strawberry bon bons	0.058	34.57899	596.18952
HersheyÕs Kisses	0.093	55.37545	595.43498
Sour Patch Kids	0.116	59.86400	516.06895
Sour Patch Tricksters	0.116	52.82595	455.39609
Root Beer Barrels	0.069	29.70369	430.48829
Sixlets	0.081	34.72200	428.66667
Smarties candy	0.116	45.99583	396.51575
Twizzlers	0.116	45.46628	391.95071
Lemonhead	0.104	39.14106	376.35631
Laffy Taffy	0.116	41.38956	356.80653
Warheads	0.116	39.01190	336.30947

Red vines	0.116	37.34852	321.97002
Starburst	0.220	67.03763	304.71649
ReeseÕs Miniatures	0.279	81.86626	293.42743
Skittles original	0.220	63.08514	286.75064
One dime	0.116	32.26109	278.11281
Skittles wildberry	0.220	55.10370	250.47134
Super Bubble	0.116	27.30386	235.37815
Lifesavers big ring gummies	0.279	52.91139	189.64656
Trolli Sour Bites	0.255	47.17323	184.99305
Nerds	0.325	55.35405	170.32015
WertherÕs Original Caramel	0.267	41.90431	156.94498
Runts	0.279	42.84914	153.58116
Tootsie Roll Snack Bars	0.325	49.65350	152.78001
Tootsie Pop	0.325	48.98265	150.71585
Kit Kat	0.511	76.76860	150.23210
Mike & Ike	0.325	46.41172	142.80528
WelchÕs Fruit Snacks	0.313	44.37552	141.77483
Milky Way Midnight	0.441	60.80070	137.87007
3 Musketeers	0.511	67.60294	132.29538
ReeseÕs Peanut Butter cup	0.651	84.18029	129.30920
Haribo Gold Bears	0.465	57.11974	122.83815
Now & Later	0.325	39.44680	121.37477
Fun Dip	0.325	39.18550	120.57079
Snickers	0.651	76.67378	117.77846
Candy Corn	0.325	38.01096	116.95681
ReeseÕs pieces	0.651	73.43499	112.80336
Milky Way	0.651	73.09956	112.28810
Junior Mints	0.511	57.21925	111.97505
ReeseÕs stuffed with pieces	0.651	72.88790	111.96298
Haribo Sour Bears	0.465	51.41243	110.56437
Peanut butter M&MÕs	0.651	71.46505	109.77734
Milk Duds	0.511	55.06407	107.75748
Peanut M&Ms	0.651	69.48379	106.73393
Caramel Apple Pops	0.325	34.51768	106.20825
Gobstopper	0.453	46.78335	103.27450
Air Heads	0.511	52.34146	102.42949
M&MÕs	0.651	66.57458	102.26510
Sugar Daddy	0.325	32.23100	99.17230
Nestle Butterfinger	0.767	70.73564	92.22378
Snickers Crisper	0.651	59.52925	91.44278
Haribo Twin Snakes	0.465	42.17877	90.70704
One quarter	0.511	46.11650	90.24757
Twix	0.906	81.64291	90.11359

Nestle Crunch	0.767	66.47068	86.66321
Tootsie Roll Juniors	0.511	43.06890	84.28356
Dots	0.511	42.27208	82.72422
100 Grand	0.860	66.97173	77.87410
Rolo	0.860	65.71629	76.41429
Charleston Chew	0.511	38.97504	76.27209
Chiclets	0.325	24.52499	75.46150
Milky Way Simply Caramel	0.860	64.35334	74.82946
Baby Ruth	0.767	56.91455	74.20410
Haribo Happy Cola	0.465	34.15896	73.46012
Swedish Fish	0.755	54.86111	72.66372
Chewey Lemonhead Fruit Mix	0.511	36.01763	70.48460
HersheyÕs Krackel	0.918	62.28448	67.84802
Almond Joy	0.767	50.34755	65.64217
HersheyÕs Special Dark	0.918	59.23612	64.52737
HersheyÕs Milk Chocolate	0.918	56.49050	61.53649
Payday	0.767	46.29660	60.36062
Mr Good Bar	0.918	54.52645	59.39701
Whoppers	0.848	49.52411	58.40108
Mounds	0.860	47.82975	55.61599
Jawbusters	0.511	28.12744	55.04391
Pop Rocks	0.837	41.26551	49.30169
Boston Baked Beans	0.511	23.41782	45.82745
Sugar Babies	0.767	33.43755	43.59524
Nestle Smarties	0.976	37.88719	38.81884
Ring pop	0.965	35.29076	36.57073
Nik L Nip	0.976	22.44534	22.99728

tootsie roll midgies

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

```
candy.pricesorted <- candy %>% arrange(desc(pricepercent))
candy.pricesorted[1:5,] %>% arrange(winpercent)
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Nik L Nip	0	1	0	0	0
Ring pop	0	1	0	0	0
Nestle Smarties	1	0	0	0	0
HersheyÕs Milk Chocolate	1	0	0	0	0
HersheyÕs Krackel	1	0	0	0	0

```
                        crispedricewafer hard bar pluribus sugarpercent
Nik L Nip                             0    0   0        1        0.197
Ring pop                              0    1   0        0        0.732
Nestle Smarties                       0    0   0        1        0.267
HersheyÕs Milk Chocolate              0    0   1        0        0.430
HersheyÕs Krackel                     1    0   1        0        0.430
                        pricepercent winpercent
Nik L Nip                      0.976   22.44534
Ring pop                       0.965   35.29076
Nestle Smarties                0.976   37.88719
HersheyÕs Milk Chocolate       0.918   56.49050
HersheyÕs Krackel              0.918   62.28448
```

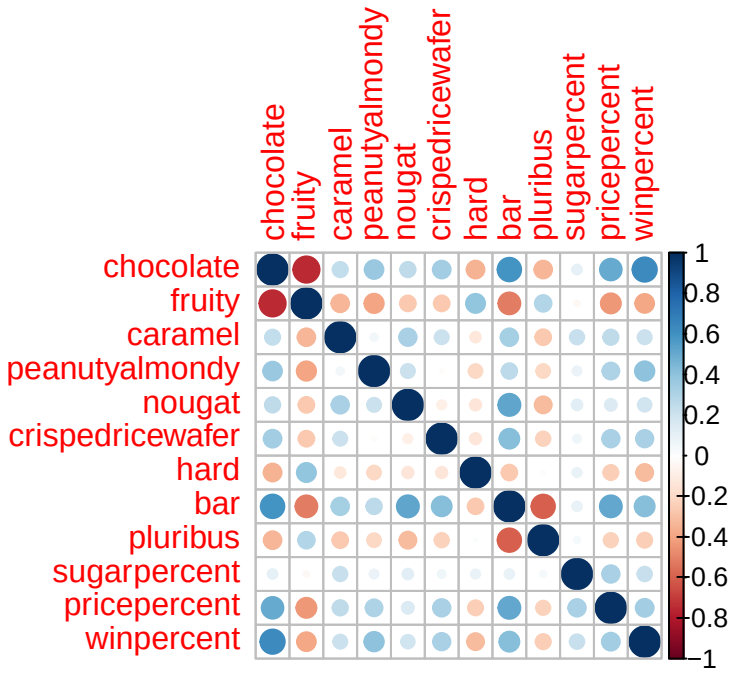
Nik L Nip is least popular of top 5 most expensive candies

5. Exploring the correlation structure

```
#install.packages("corrplot")
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



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

chocolate and fruity (not really many fruity chocolate candies at all)

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

chocolate and bar and chocolate and winpercent (lots of chocolate bars and people like chocolate)

6. PCA: Principal Component Analysis

the main function that always there for us is `prcomp()`. it has an important argument that is set to `scale=FALSE`

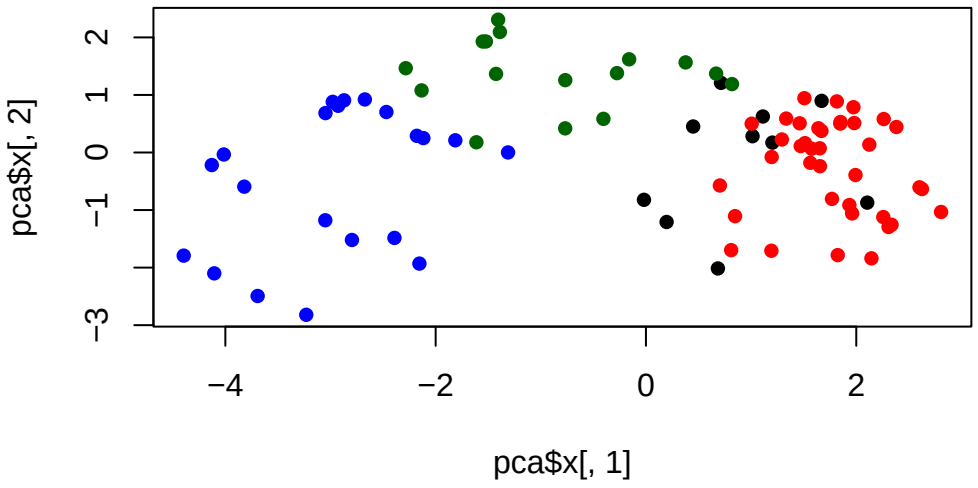
```
#need to scale because we saw that the winpercent values are on 1-100 while others are 0-1
pca <- prcomp(candy,scale=TRUE)
summary(pca)
```

```
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
```

my PCA plot (a.k.a.) PC1 vs PC2 score plot.

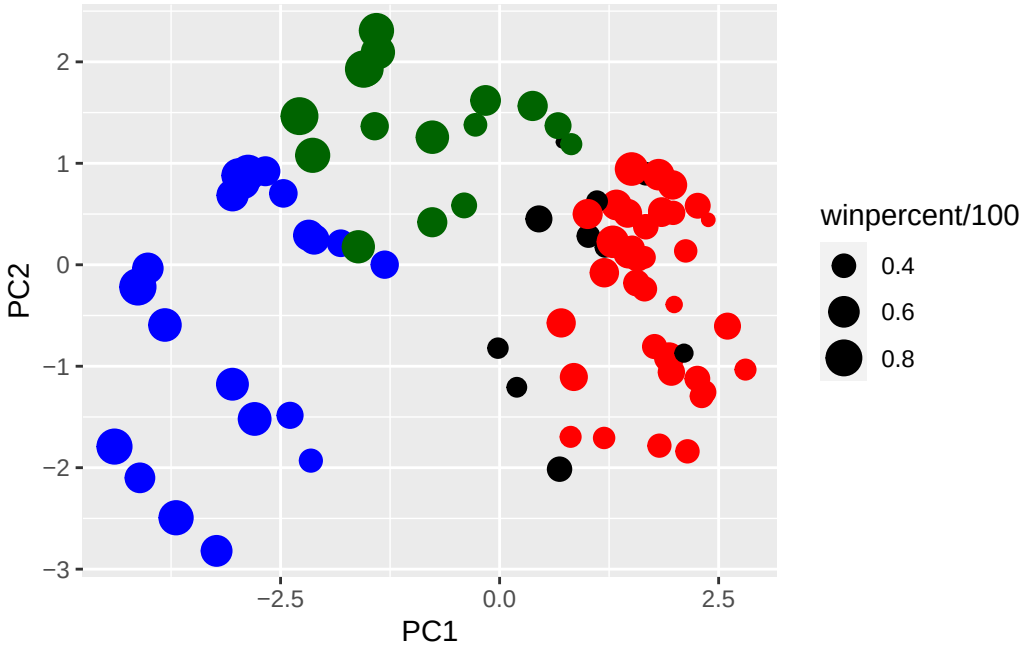
```
plot(pca$x[,1],pca$x[,2],col=my_cols,pch=16)
```



ggplot to make things prettier

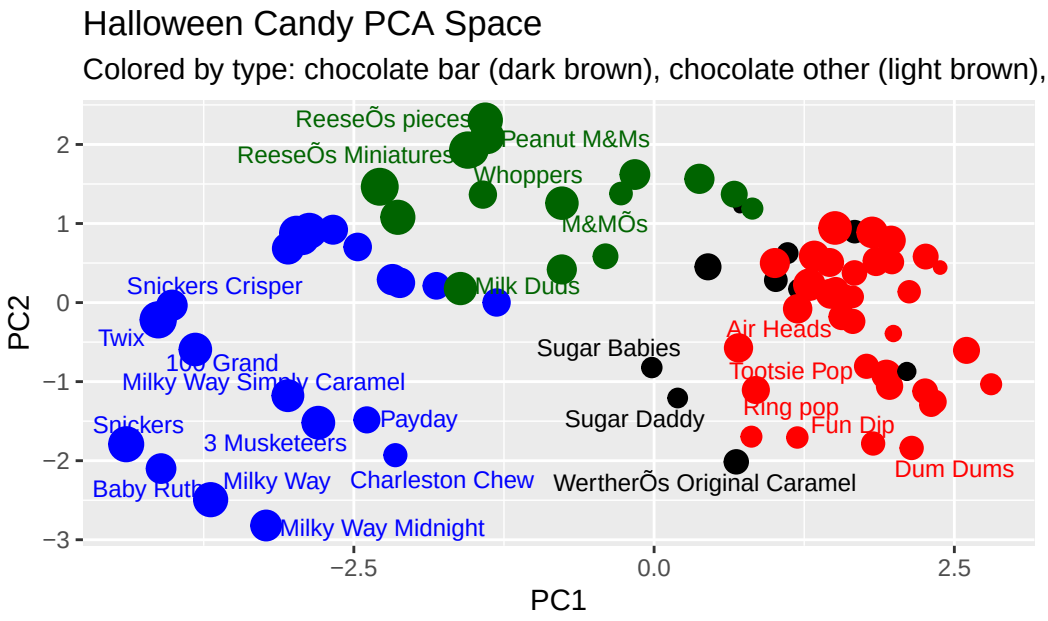
```
#make a data frame for ggplotting
my_data<-cbind(candy,pca$x[,1:3])
```

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

p
```




```
#use ggrepel to add labels
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
```




```
#interactive plot with plotly
library(plotly)
```

```
Attaching package: 'plotly'
```

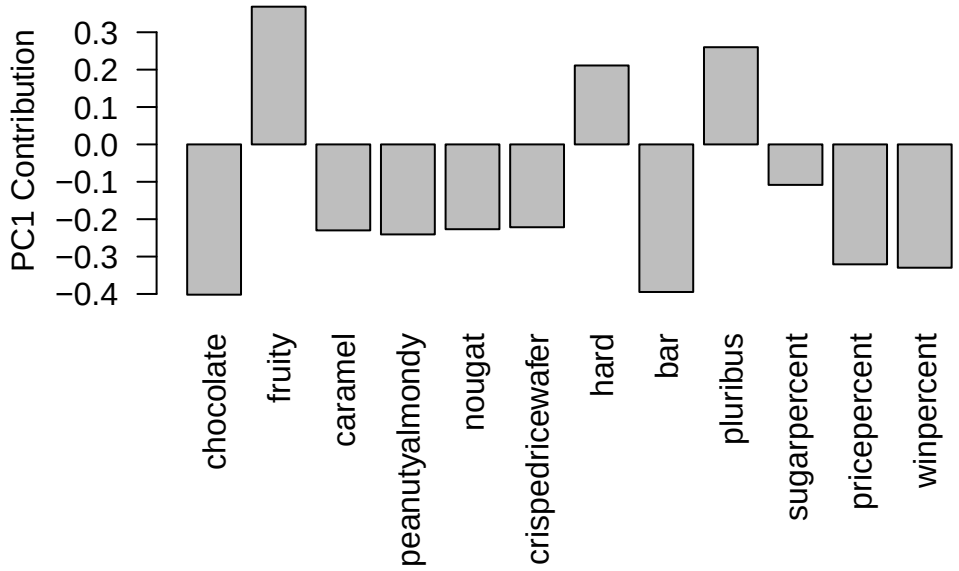
```
The following object is masked from 'package:ggplot2':

    last_plot
```

```
The following object is masked from 'package:stats':

    filter
```

```
The following object is masked from 'package:graphics':

    layout
```

```
#ggplotly(p)
#hashtagging plotly out for pdf purposes
```

```r
#let's see each candy type contribution to 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, pluribus. negative is chocolate, bar, priceprecent, winpercent, etc. These make sense because we see these based off of where they lie on PC1 in the graph.