Halloween MiniProject

Diana Furlan

2024-10-28

```
url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/ca

candy_file <- read.csv(url)
candy = read.csv(url, 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?

```
candies <- nrow(candy)
candies
```

```
[1] 85
```

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

```
fruity <- sum(candy$fruity == 1)
fruity
```

```
[1] 38
```

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

```
candy["Almond Joy", ]$winpercent
```

```
[1] 50.34755
```

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
```

There is a useful'skim()'function in the skimr package that can help give you a quick overview of a given dataset

```
#install.packages("skimr")
library("skimr")
#or just extract a part from the library
#skimr::skim(candy)
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? sugarpercent, pricepercent and winpercent are on a non binary scale.

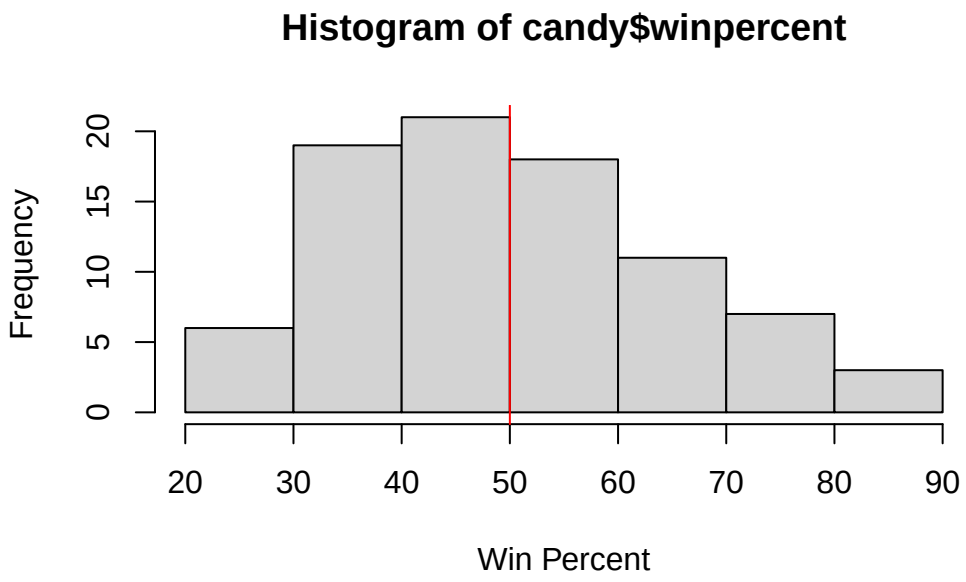
Q7. What do you think a zero and one represent for the candy$chocolate column? 1 = T for chocolate and 0 = F for chocolate

---

Q8. Plot a histogram of winpercent values

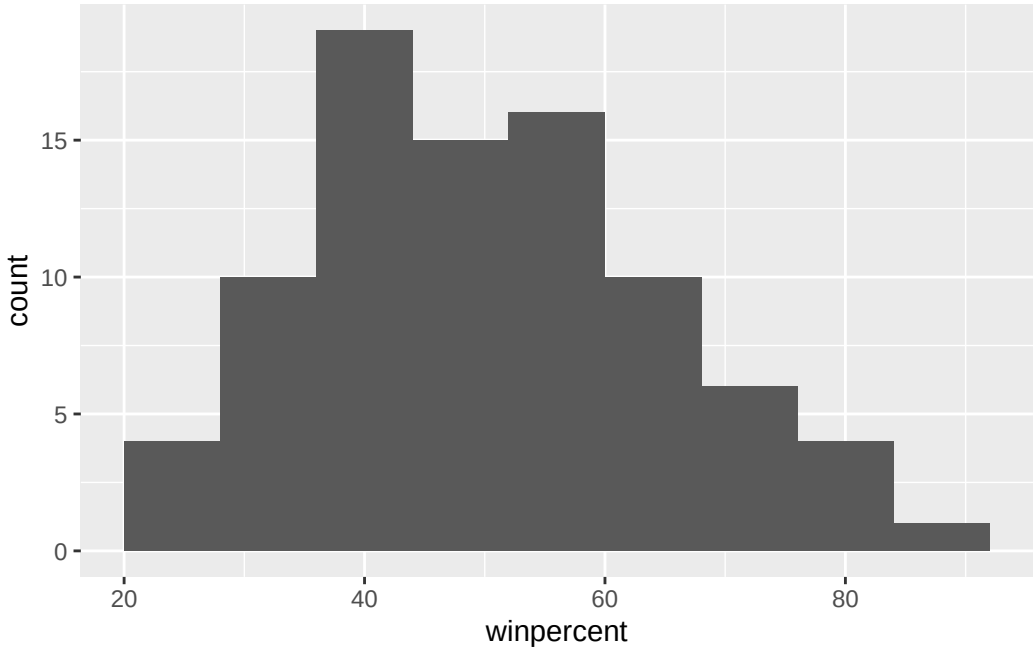
```
#can add color with 'col'- how to change the title?

hist(candy$winpercent, xlab = "Win Percent")
  abline(v = 50, col = "red")
```




```
library(ggplot2)

ggplot(candy) +
  aes(winpercent) +
  geom_histogram(binwidth = 8)
```



Q9. Is the distribution of winpercent values symmetrical? slighly right-skewed

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

```
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.14   47.83   50.32   59.86   84.18 
```

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

```
summary(candy[as.logical(candy$chocolate),]$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  34.72   50.35   60.80   60.92   70.74   84.18 
```

```
#installed.packages("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
```

```
fruit.candy <- candy |>
   filter(fruity==1)

summary(fruit.candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.45   39.04   42.97   44.12   52.11   67.04
```

Q12. Is this difference statistically significant?

```
choc.candy <- candy |> filter(chocolate ==1)
fruit.candy <- candy |> filter(fruity == 1)
```

```
t_test_result <- t.test(choc.candy$winpercent, fruit.candy$winpercent)

t_test_result
```

```
    Welch Two Sample t-test

data:  choc.candy$winpercent and fruit.candy$winpercent
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
```

```
play <- c("d","a","c")
sort(play)
```

```
[1] "a" "c" "d"
```

```
order(play)
```

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

```
play[order(play) ]
```

```
[1] "a" "c" "d"
```

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

```
sort(c(5, 2, 10), decreasing = T)
```

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

```
head( candy[order(candy$winpercent),], 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
```

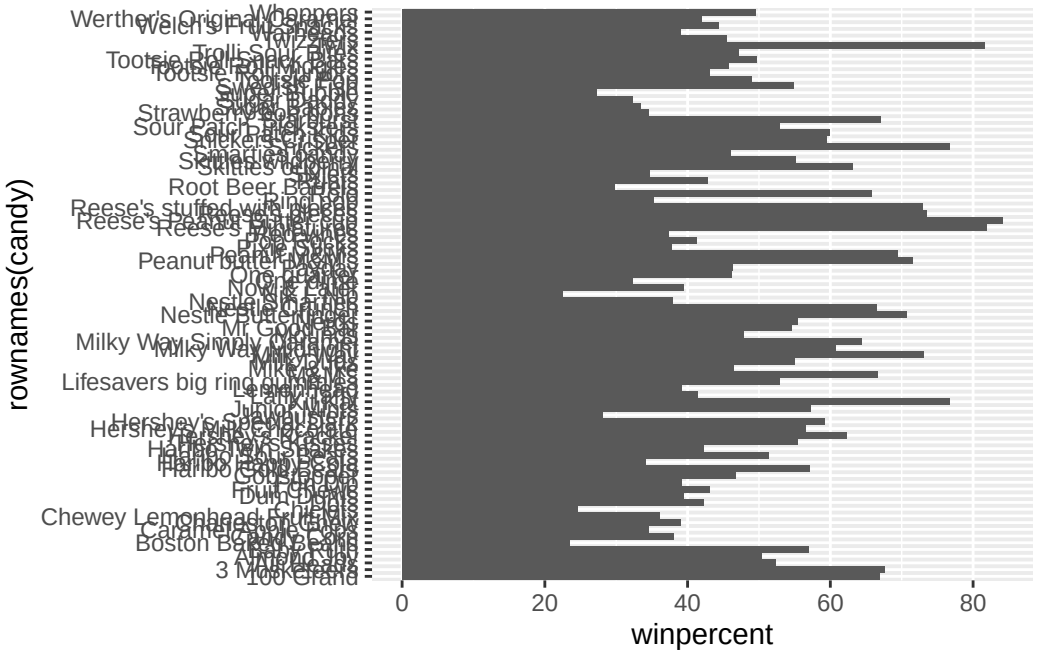
Q14. What are the top 5 all time favorite candy types out of this set?

```
#candy%>%
 # arrange(winpercent) %>% head(5)
```

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

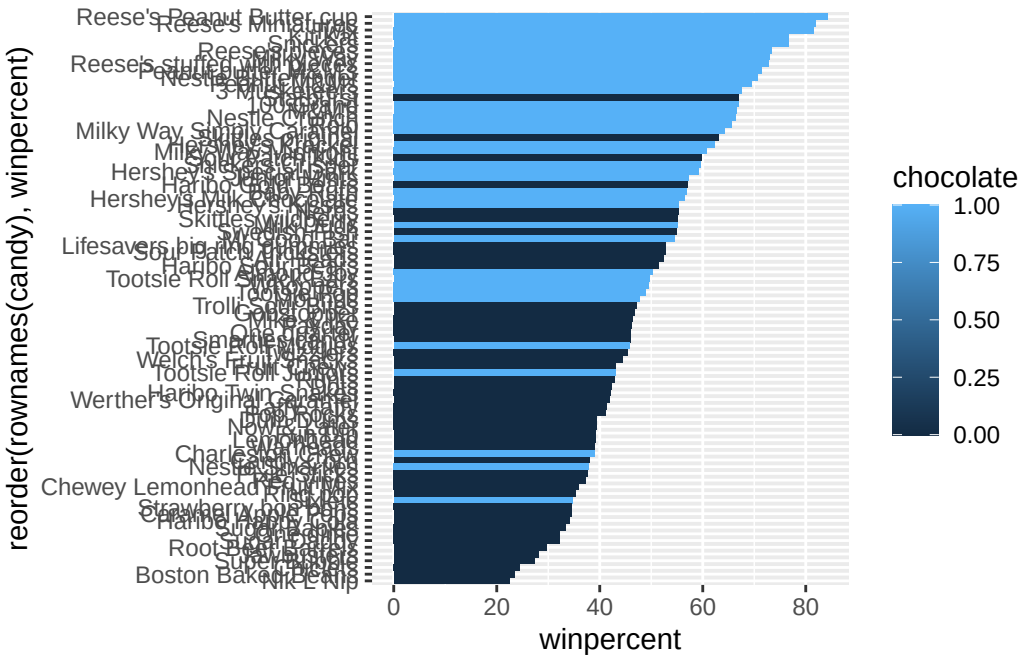
```
library(ggplot2)

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



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

```
ggplot(candy) +
  aes(x=winpercent,
      y=reorder(rownames(candy), winpercent),
      fill= chocolate) +
  geom_col()
```



More custom col skim so we can see both chocolate and bar and fruity, etc. all from the same plot

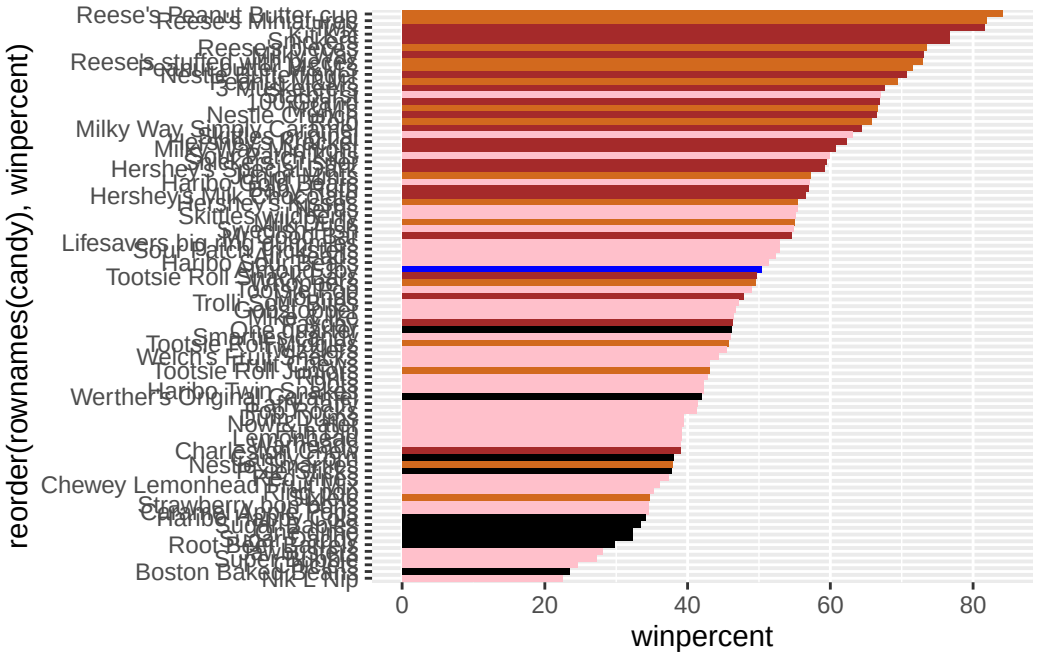
```
mycols <- rep("black", nrow(candy))
mycols[as.logical(candy$chocolate)] <- "chocolate"
mycols [as.logical(candy$bar)] <- "brown"
mycols [as.logical(candy$fruity)] <- "pink"
#use blue for my fav candy
```

```
mycols [rownames(candy)=="Almond Joy"] <- "blue"
mycols
```

```
 [1] "brown"     "brown"     "black"     "black"     "pink"      "blue"
 [7] "brown"     "black"     "black"     "pink"      "brown"     "pink"
[13] "pink"      "pink"      "pink"      "pink"      "pink"      "pink"
[19] "pink"      "black"     "pink"      "pink"      "chocolate" "brown"
[25] "brown"     "brown"     "pink"      "chocolate" "brown"     "pink"
[31] "pink"      "pink"      "chocolate" "chocolate" "pink"      "chocolate"
[37] "brown"     "brown"     "brown"     "brown"     "brown"     "pink"
[43] "brown"     "brown"     "pink"      "pink"      "brown"     "chocolate"
[49] "black"     "pink"      "pink"      "chocolate" "chocolate" "chocolate"
[55] "chocolate" "pink"      "chocolate" "black"     "pink"      "chocolate"
```

```
[61] "pink"      "pink"      "chocolate" "pink"      "brown"     "brown"
[67] "pink"      "pink"      "pink"      "pink"      "black"     "black"
[73] "pink"      "pink"      "pink"      "chocolate" "chocolate" "brown"
[79] "pink"      "brown"     "pink"      "pink"      "pink"      "black"
[85] "chocolate"
```

```
#placeholder
ggplot(candy) +
  aes(x=winpercent,
      y=reorder(rownames(candy), winpercent),) +
  geom_col(fill =mycols)
```



Q17. What is the worst ranked chocolate candy? Nik L Nip

Q18. What is the best ranked fruity candy? Starbust

```
mycols[as.logical(candy$fruity)]<- "red"
mycols
```

```
 [1] "brown"     "brown"     "black"     "black"     "red"       "blue"
 [7] "brown"     "black"     "black"     "red"       "brown"     "red"
[13] "red"       "red"       "red"       "red"       "red"       "red"
```

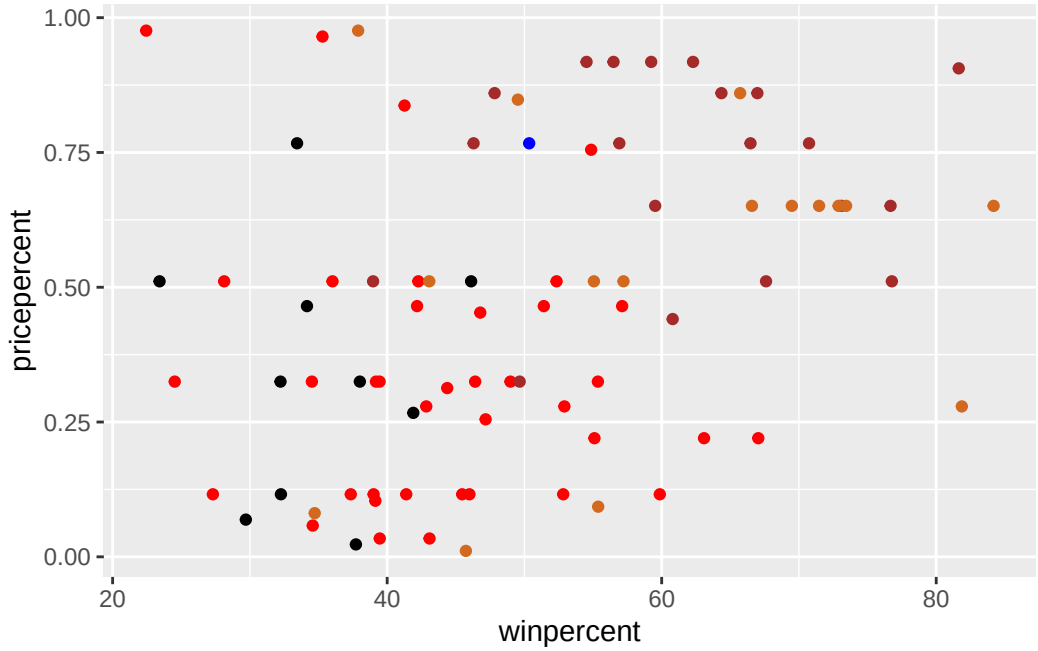
```
 [19] "red"       "black"     "red"       "red"       "chocolate" "brown"
 [25] "brown"     "brown"     "red"       "chocolate" "brown"     "red"
 [31] "red"       "red"       "chocolate" "chocolate" "red"       "chocolate"
 [37] "brown"     "brown"     "brown"     "brown"     "brown"     "red"
 [43] "brown"     "brown"     "red"       "red"       "brown"     "chocolate"
 [49] "black"     "red"       "red"       "chocolate" "chocolate" "chocolate"
 [55] "chocolate" "red"       "chocolate" "black"     "red"       "chocolate"
 [61] "red"       "red"       "chocolate" "red"       "brown"     "brown"
 [67] "red"       "red"       "red"       "red"       "black"     "black"
 [73] "red"       "red"       "red"       "chocolate" "chocolate" "brown"
 [79] "red"       "brown"     "red"       "red"       "red"       "black"
 [85] "chocolate"
```

```
#install.packages("ggrepel")
```

```
#library(ggrepel)
#ggplot (candy) +
 # aes(winpercent,pricepercent, label=rownames(candy)) +
  #geom_point(col=mycols) +
  #geom_label(col=mycols) +
  #geom_text_repel(max.overlaps = 8, col=mycols, )
```

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?

```
ggplot (candy)+
  aes(winpercent,pricepercent)+
  geom_point(col=mycols)
```



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

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

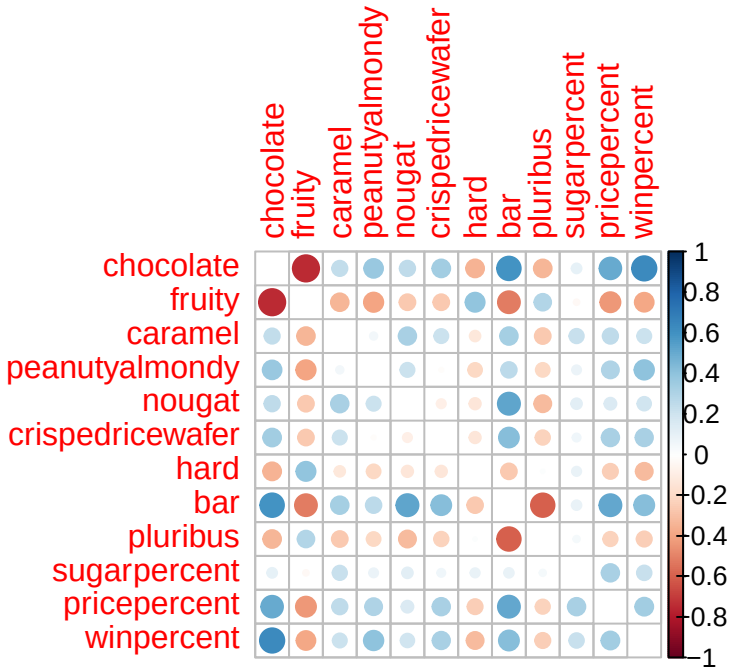
```
                      pricepercent winpercent
Nik L Nip                    0.976   22.44534
Nestle Smarties              0.976   37.88719
Ring pop                     0.965   35.29076
Hershey's Krackel            0.918   62.28448
Hershey's Milk Chocolate     0.918   56.49050
```

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

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
cij <- cor(candy)
corrplot (cij, diag=F)
```



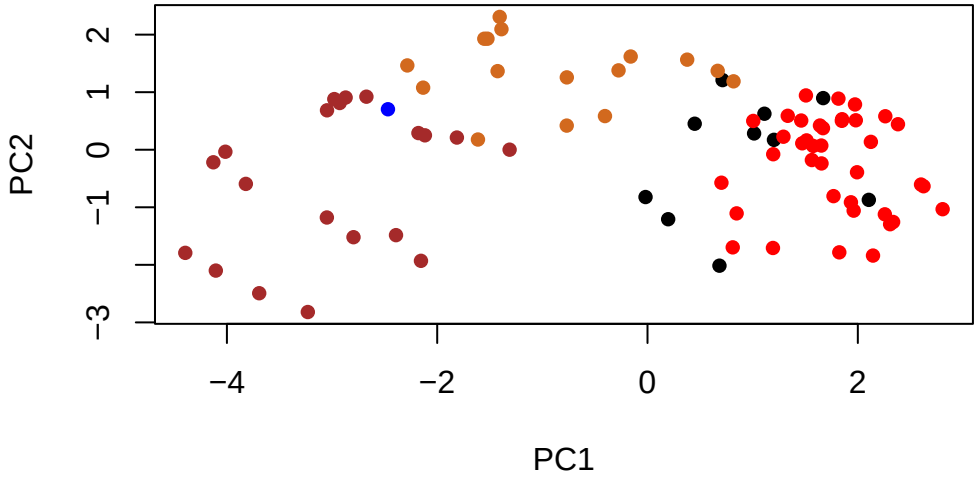
Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)? chocolate and gruit

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

```
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
```

```
plot(pca$x[,1], pca$x[,2], col=mycols, pch=16, xlab = "PC1", ylab = "PC2")
```

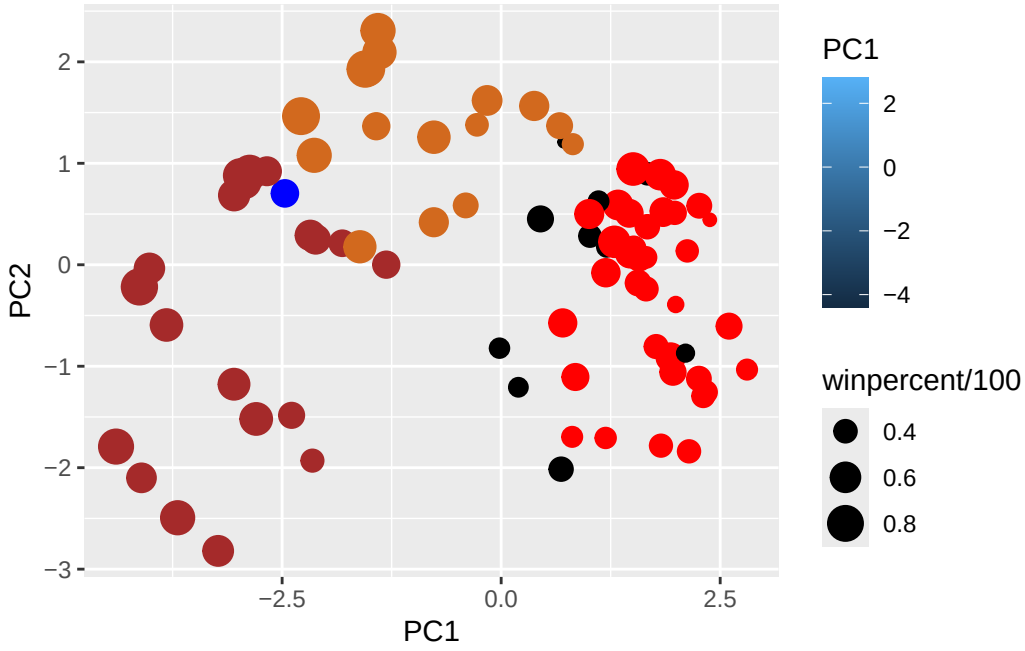


how do the original variables cols contribute to the new pca. PC1:

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

```
p <- ggplot(loadings) +
  aes(x= PC1, y= PC2,
      text=rownames(loadings), fill = PC1, size= winpercent/100) +
  geom_point(col=mycols)

  p
```



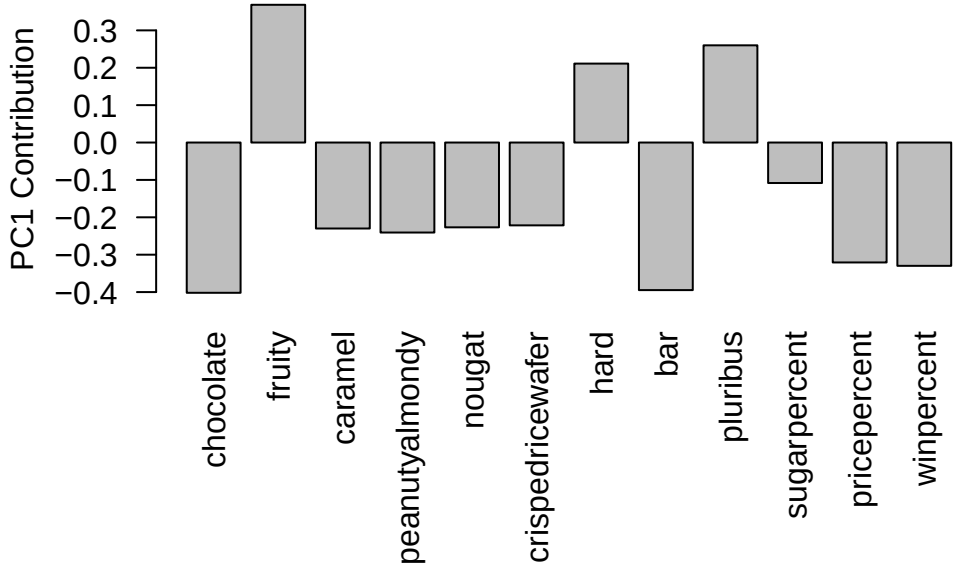

```
#install.packages("ggrepel")
#library(ggrepel)

#q <- ggplot(loadings) +
 # aes(x = PC1, y = PC2, text = rownames(loadings), fill = PC1, size = winpercent / 100) +
  #geom_point(col=mycols) +
  #geom_text_repel(aes(label = rownames(loadings)), size = 3.3, col = mycols, max.overlaps =
  #theme(legend.position = "none") +
  #labs(title = "Halloween Candy PCA Space", subtitle = "Colored by type: chocolate bar (dar
   # caption = "Data from 538")
#q
```

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

#ggplotly(q)
```

```
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?No,I think it should be reveresed