HalloweenMiniProject

Dani Weatherwax (PID:A17856408)

Quarto

```
library(readr)
candy_data <- read_csv("~/Downloads/candy-data.csv")
```

```
Rows: 85 Columns: 13
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): competitorname
dbl (12): chocolate, fruity, caramel, peanutyalmondy, nougat, crispedricewaf...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
candy_file <- "~/Downloads/candy-data.csv"

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
```

```
nrow(candy)
```

```
[1] 85
```

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

```
[1] 38
```

Q1. There are 85 different candy types in this dataset. Q2. There are 38 fruity candies in this dataset.

```
candy["Reese's Peanut Butter cup", ]$winpercent
```

```
[1] 84.18029
```

Q3. My favorite candy in the dataset is reese's peanut butter cups, and it's winpercent is 84.18029%.

Q4. Kit Kat's winpercent is 76.7686%

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

```
[1] 76.7686
```

Q5. The winpercent for Tootsie Roll Snack bars is 49.6535%

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

```
[1] 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

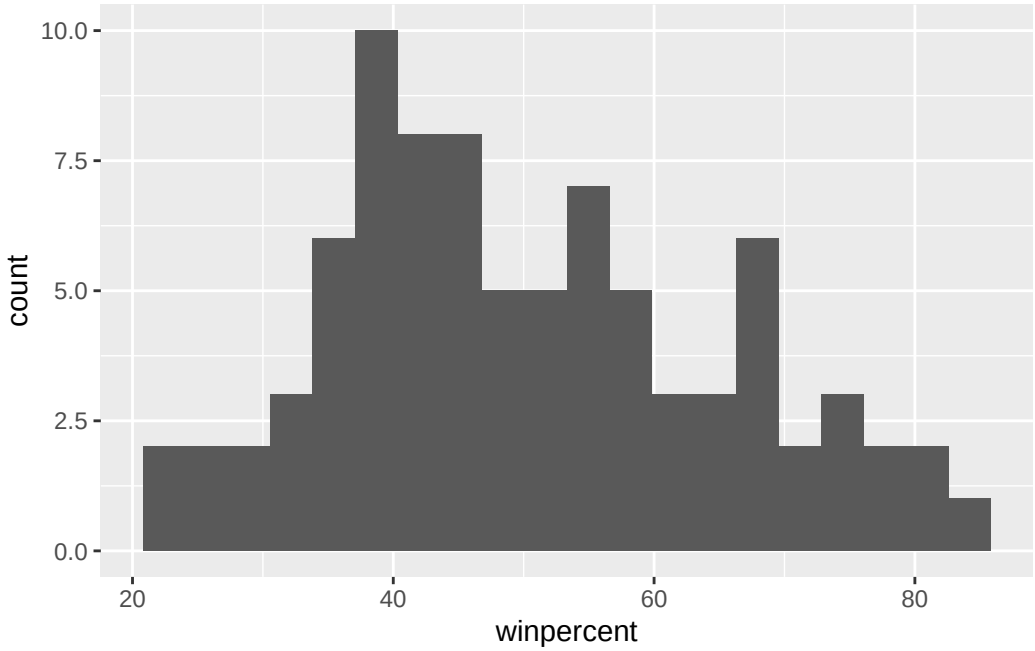
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.00	1.00	
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00	
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
nougat	0	1	0.08	0.28	0.00	0.00	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. The winpercent variable seems to be much larger than the other rows.

Q7. The 0 would represent the candy lacking chocolate, while a 1 would represent the candy having chocolate.

Q9.The distribution of win percent isn't perfectly symmetrical.

```
library(ggplot2)
ggplot(candy, aes(winpercent))+
  geom_histogram(bins=20)
```



Q10. The median is less than 50% (47.83%), so thus the center of distribution must be below 50%.

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

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

1. Find all chocolate candy in the dataset

2. Find their winpercent values

3. Calculate the mean of these values

4-6. Do the same for fruity candy

7. Compare mean winpercents for chocolate vs. fruity

```
choc.inds <- candy$chocolate == 1
choc.win <- candy[choc.inds,]$winpercent
mean(choc.win)
```

```
[1] 60.92153
```

```
frt.inds <- candy$fruity == 1
frt.win <- candy[frt.inds,]$winpercent
mean(frt.win)
```

```
[1] 44.11974
```

Q11. Chocolate candy is ranked higher than fruity candy.

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

```
    Welch Two Sample t-test

data:  choc.win and frt.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
```

The difference in ranking between the two candies is statistically significant.

> Q13. The five least liked candies are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

```
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
```

```
library(dplyr)
candy %>%
  arrange(-winpercent) %>%
head(5)
```

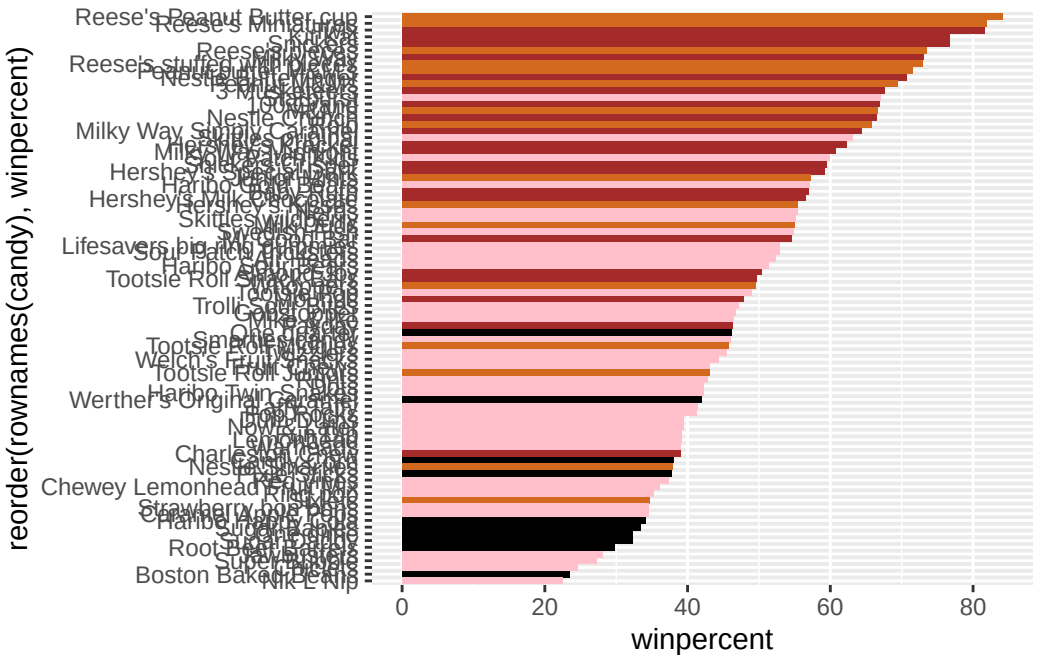
```
                         chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           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
```

Q14. The top five all time candy types are Reese's PB cups, Minis, Twix, Kit Kats, and Snickers.

```
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"
```

```
library(ggplot2)

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

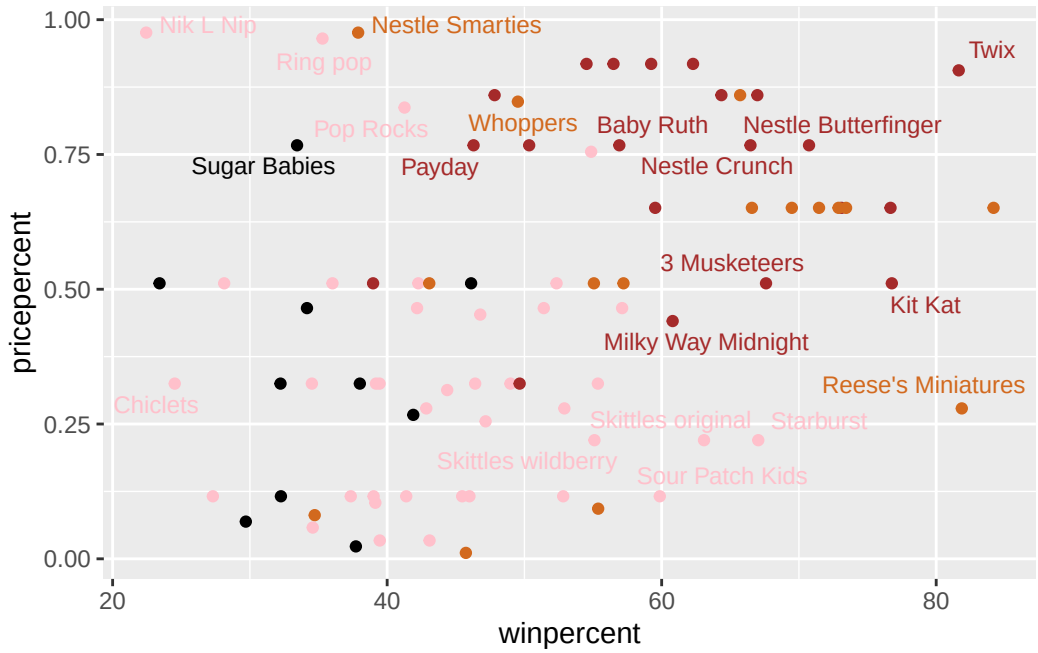


Q17. The worst ranked chocolate candy is sixlets.

Q18. The best ranked fruity candy is starburst.

```
library(ggrepel)
# How about a plot of price vs win
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. The biggest bang for your buck is Tootsie Roll Midgies!!

```
candy |>
mutate(bang_for_buck = winpercent / pricepercent) |>
arrange(desc(bang_for_buck)) |>
slice(1)
```

```
                    chocolate fruity caramel peanutyalmondy nougat
Tootsie Roll Midgies         1      0       0              0      0
                    crispedricewafer hard bar pluribus sugarpercent
Tootsie Roll Midgies                0    0   0        1        0.174
                    pricepercent winpercent bang_for_buck
Tootsie Roll Midgies        0.011   45.73675      4157.886
```

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

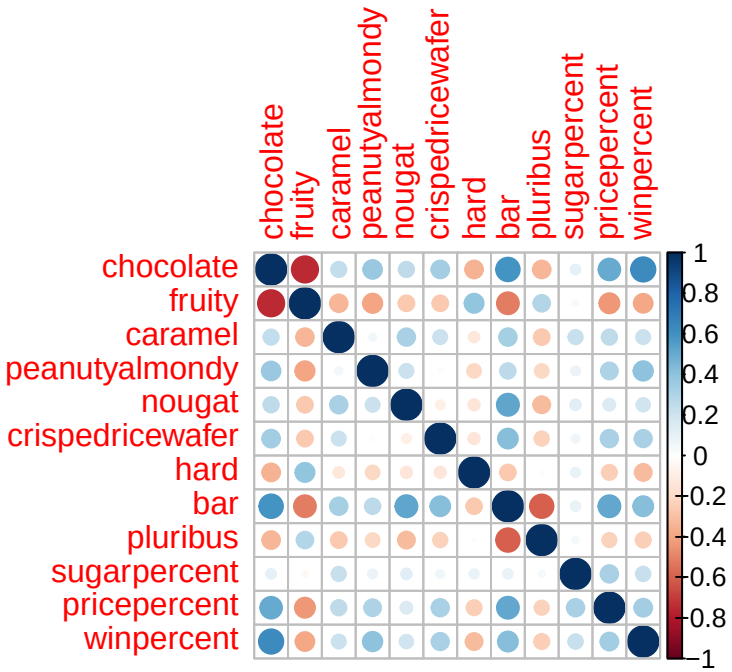
```
                      pricepercent winpercent
Nik L Nip                    0.976   22.44534
Nestle Smarties              0.976   37.88719
Ring pop                     0.965   35.29076
Hershey's Krackel            0.918   62.28448
Hershey's Milk Chocolate     0.918   56.49050
```

Q20. The most expensive are Nik L Nip, Smarties, Ring Pops, Krackel, and Hershey's Milk Chocolate, with Nik L Nip having the lowest win percentage.

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



Q22. Fruity and chocolate are anti-correlated.

```
which(cij == min(cij), arr.ind = TRUE)
```

```
          row col
fruity      2   1
chocolate   1   2
```

Q23. The two most positively correlated variables are chocolate and winpercent.

```r
cij_no_diag <- cij
diag(cij_no_diag) <- NA
which(cij_no_diag == max(cij_no_diag, na.rm = TRUE), arr.ind = TRUE)
```

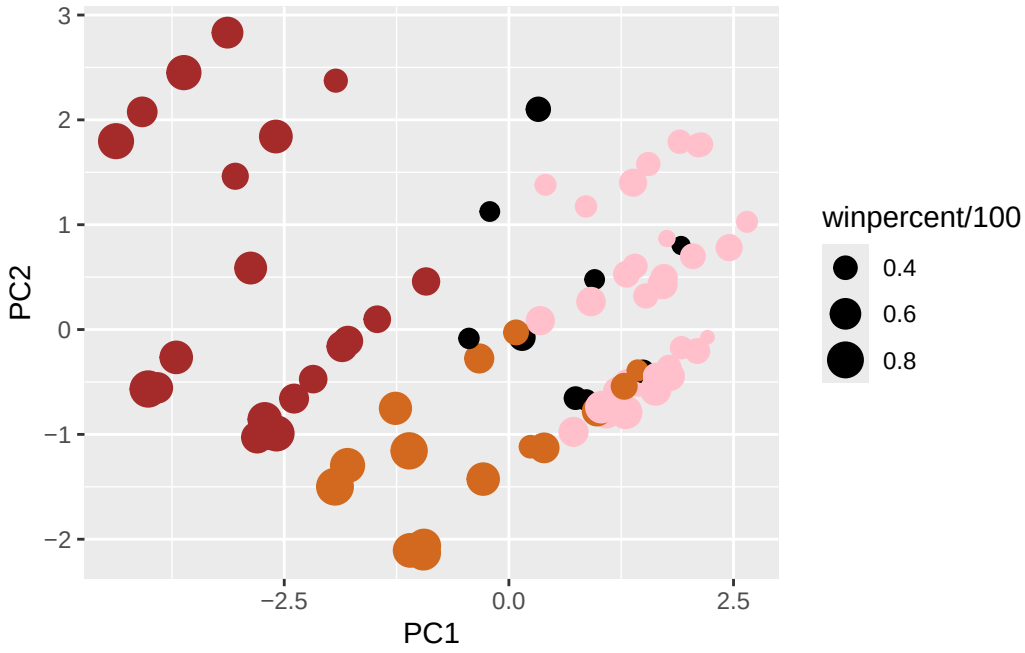
```
           row col
winpercent  12   1
chocolate    1  12
```

```r
candy <- candy[, -1]
pca <- prcomp(candy, scale = T)
summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     1.9200 1.1143 1.1085 1.0751 0.95010 0.81815 0.81352
Proportion of Variance 0.3351 0.1129 0.1117 0.1051 0.08206 0.06085 0.06016
Cumulative Proportion  0.3351 0.4480 0.5597 0.6648 0.74685 0.80770 0.86787
                           PC8     PC9    PC10    PC11
Standard deviation     0.68950 0.64410 0.60875 0.43887
Proportion of Variance 0.04322 0.03772 0.03369 0.01751
Cumulative Proportion  0.91109 0.94880 0.98249 1.00000
```

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

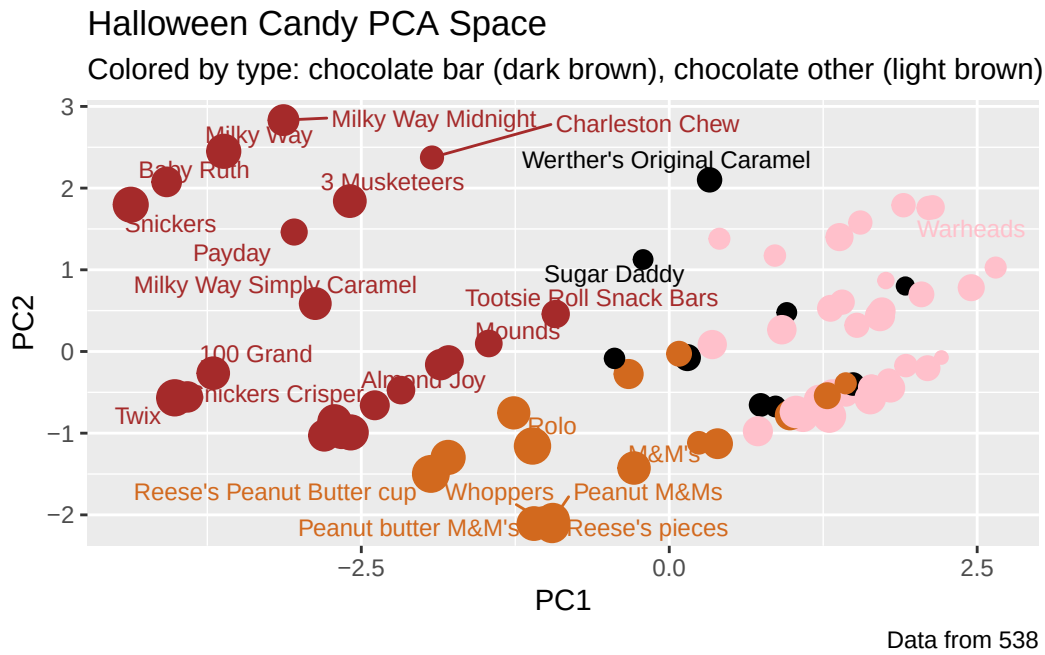
```r
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)
p
```



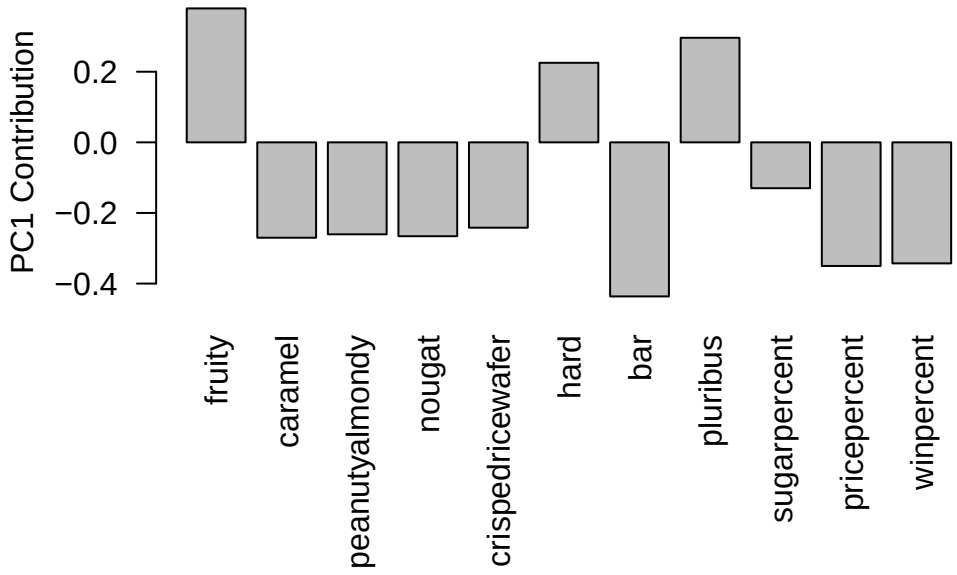

```
library(ggrepel)

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




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



Q24. The variables picked up strongly by PC1 are hard and pluribus. This makes sense, as many fruity candies come in boxes and are hard (e.g. skittles)