class10

Steven Gan

2022-10-28

Table of contents

1. Importing candy data

```
candy <- read.csv(file = "https://raw.githubusercontent.com/fivethirtyeight/data/master/ca
                  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
```

Q2:

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

```
[1] 38
```

2. What is your favorate candy?

```
candy["Twix", ]$winpercent
```

```
[1] 81.64291
```

Q3:

```
print("Milky Way")
```

```
[1] "Milky Way"
```

```
candy["Milky Way", ]$winpercent
```

```
[1] 73.09956
```

Q4:

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

```
[1] 76.7686
```

Q5:

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

```
[1] 49.6535
```

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

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

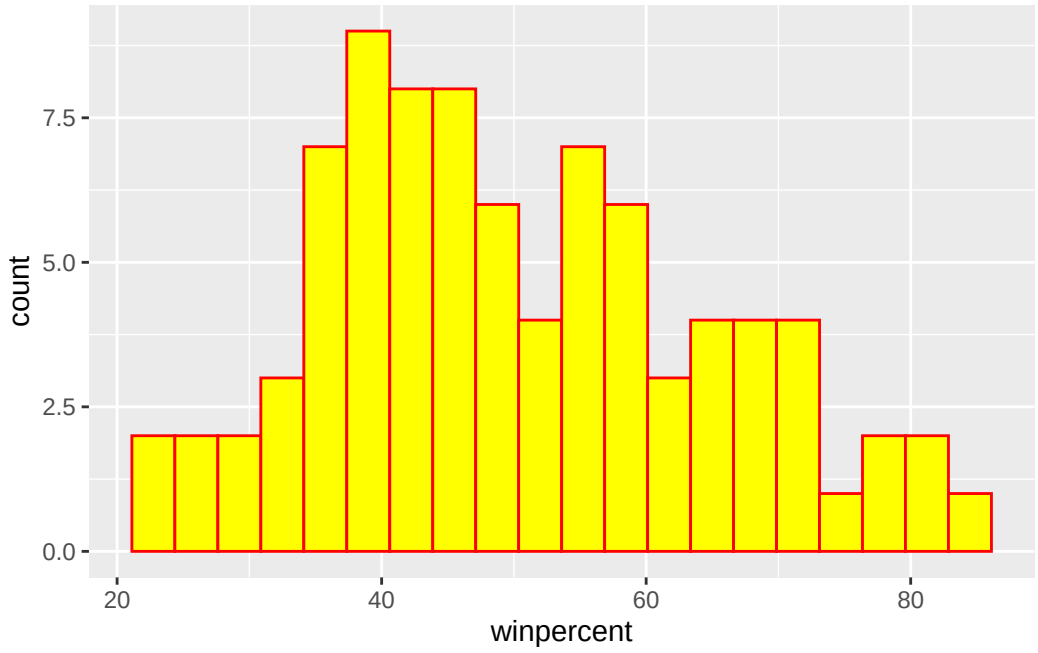
Q6: The variable "winpercent" is different in scale, which is from 0 to 100.

Q7: 1 and 0 means whether that type of candy contains chocolate or not.

Q8:

```
library(ggplot2)

ggplot(candy, aes(x = winpercent)) +
  geom_histogram(bins = 20, fill = "yellow", col = "red")
```



Q9: The distribution of winpercent values is not completely symmetrical.

Q10: Below

Q11:

```
chocoWinper <- candy[as.logical(candy$chocolate), ]$winpercent
fruWinper <- candy[as.logical(candy$fruity), ]$winpercent

if (mean(chocoWinper) > mean(fruWinper))
  {print("Higher")} else {print("Lowwer")}
```

```
[1] "Higher"
```

Q12:

```
t.test(chocoWinper, fruWinper)
```

```

    Welch Two Sample t-test

data:  chocoWinper and fruWinper
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
```

Different statistically significant

3. Overall Candy Rankings

Q13:

```
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) %>%
  row.names()
```

```
[1] "Nik L Nip"           "Boston Baked Beans" "Chiclets"
[4] "Super Bubble"        "Jawbusters"
```

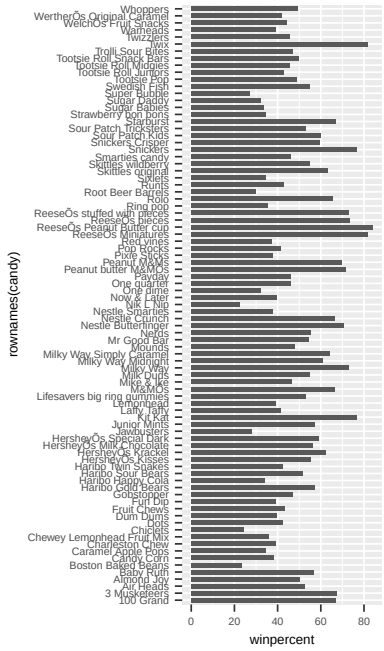
Q14:

```
candy %>%
  arrange(desc(winpercent)) %>%
  head(5) %>%
  row.names()
```

```
[1] "ReeseÕs Peanut Butter cup" "ReeseÕs Miniatures"
[3] "Twix"                      "Kit Kat"
[5] "Snickers"
```

Q15:

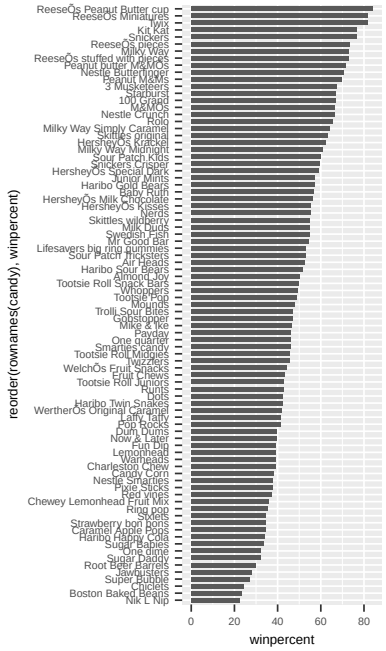
```
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col(width = 0.7) +
  theme(text = element_text(size = 5), element_line(size = 0.3),
        aspect.ratio = 3)
```




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

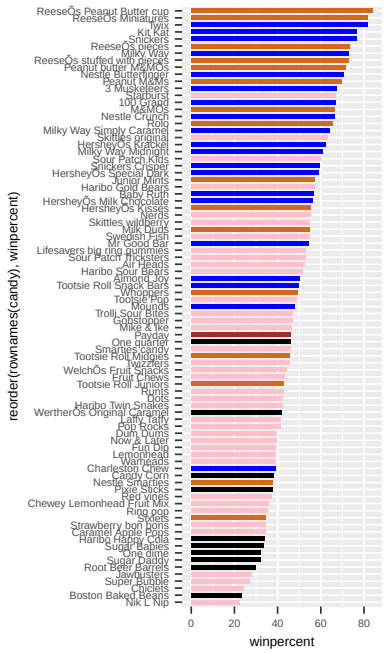
Q16:

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col(width = 0.7) +
  theme(text = element_text(size = 5), element_line(size = 0.3),
        aspect.ratio = 3)
```




```
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$chocolate) & as.logical(candy$bar)] = "blue"
my_cols[as.logical(candy$fruity)] = "pink"
```

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(width = 0.7, fill = my_cols) +
  theme(text = element_text(size = 5), element_line(size = 0.3),
        aspect.ratio = 3)
```




```
# gsave("ChocoBarFru.png")
```

Q17: Sixlets

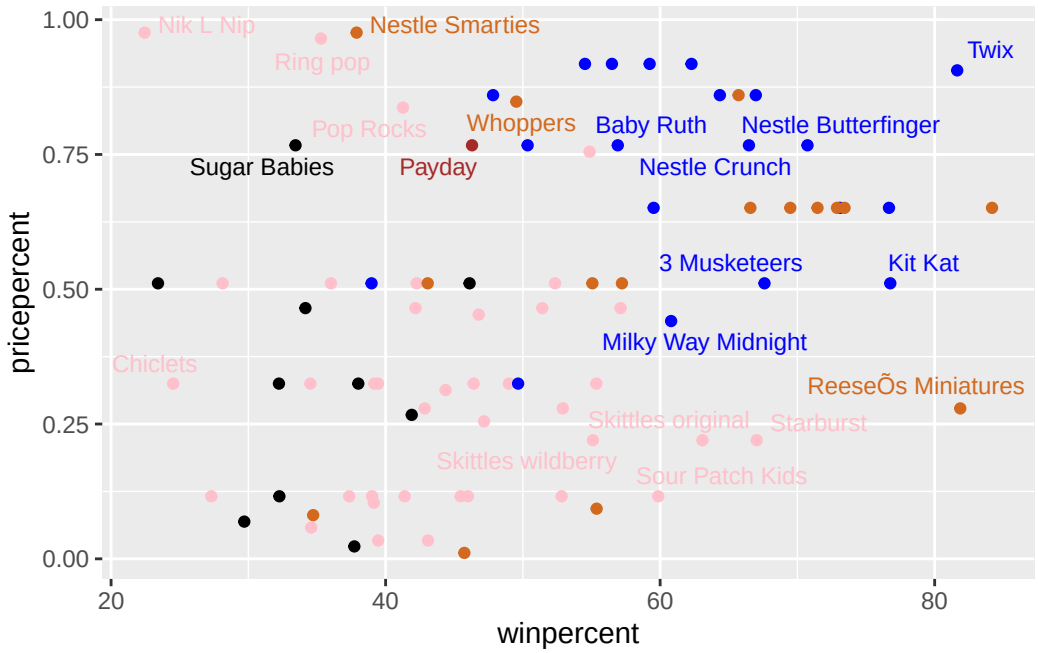
Q18: Starburst

4. Taking a look at pricepercent

```
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: Probably Sour Patch Kids, Starburst, ReeseÕs Miniatures

Q20:

```
candy %>%
  arrange(desc(pricepercent)) %>%
  head(5) %>%
  row.names()
```

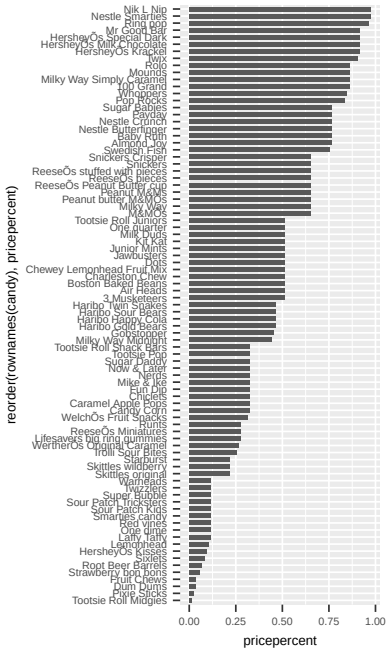
```
[1] "Nik L Nip"                "Nestle Smarties"
[3] "Ring pop"                 "HersheyÕs Krackel"
[5] "HersheyÕs Milk Chocolate"
```

```
candy %>%
  arrange(desc(pricepercent)) %>%
  head(5) %>%
  arrange(winpercent) %>%
  head(1) %>%
  row.names()
```

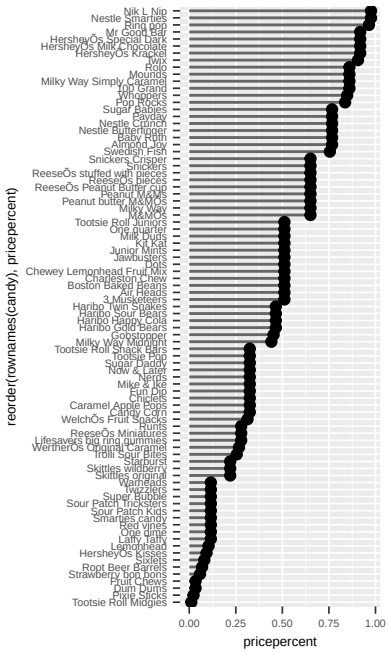
```
[1] "Nik L Nip"
```

Q21:

```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_col(width = 0.7) +
  theme(text = element_text(size = 5), element_line(size = 0.3),
        aspect.ratio = 3)
```




```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col = "gray40") +
  geom_point() +
  theme(text = element_text(size = 5), element_line(size = 0.3),
        aspect.ratio = 3)
```

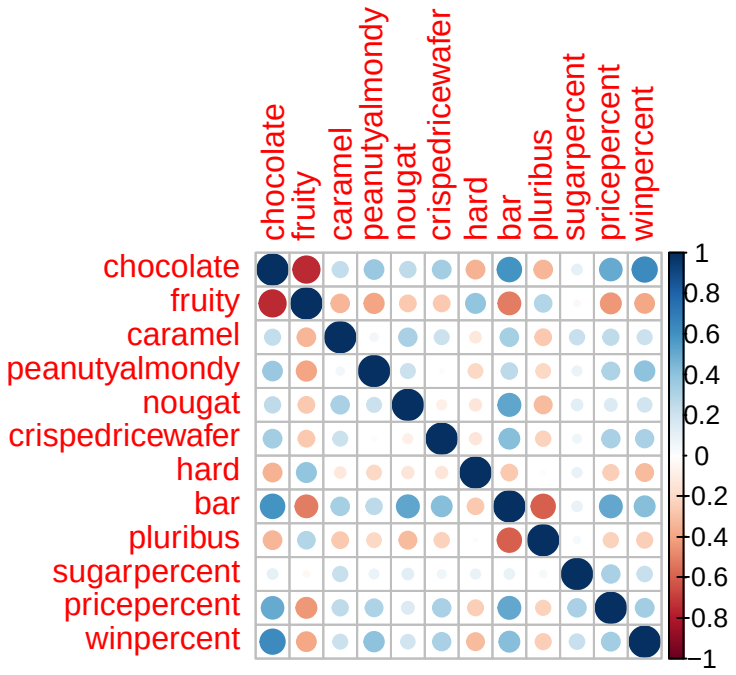


5 Exploring the correlation structure

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



Q22: Chocolate and fruity are anti-correlated

Q23: Chocolate and Winpercent are correlated

6. Principal Component Analysis

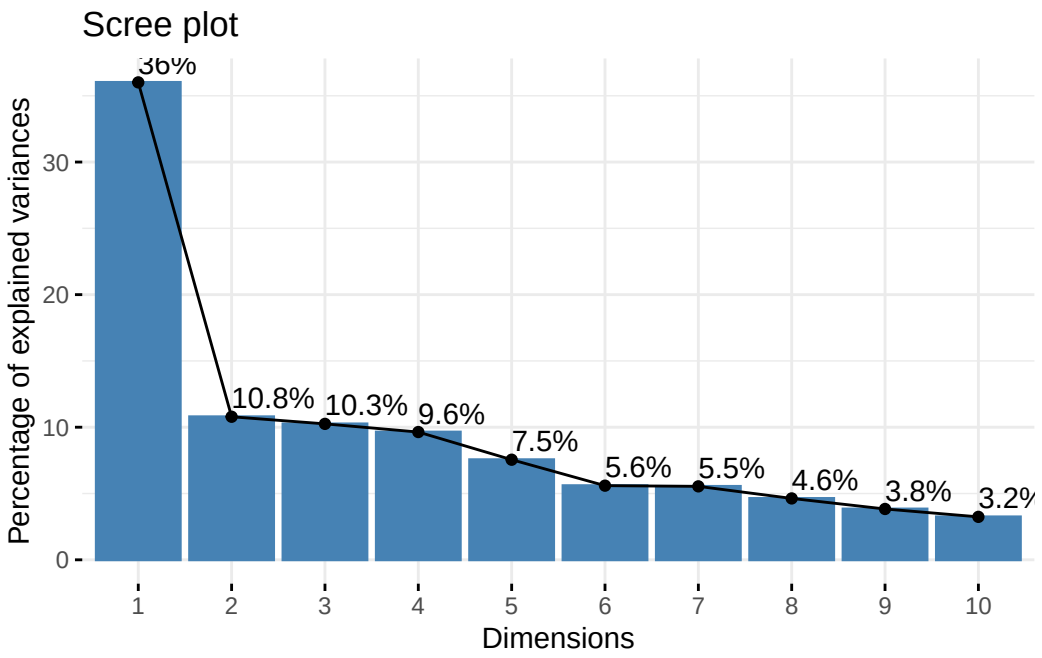
```
pca <- prcomp(candy, scale = T)
summary(pca)
```

```
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
```

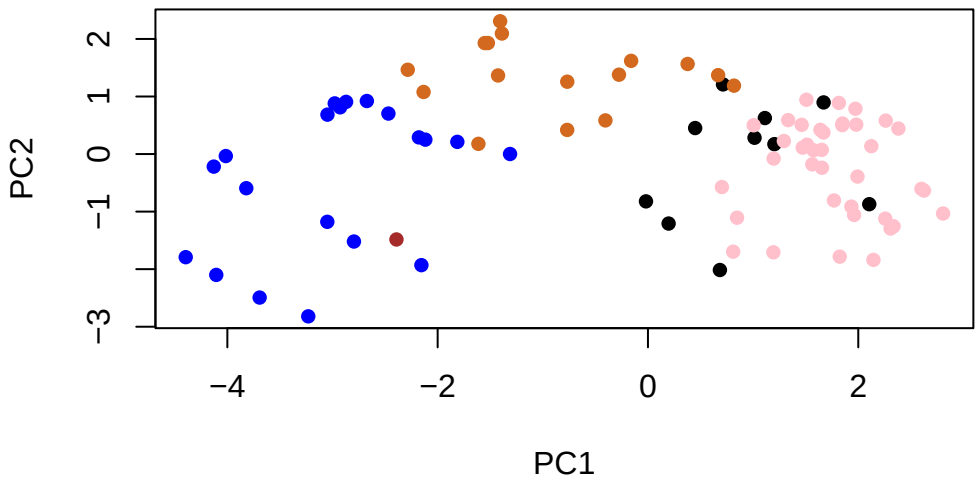
```
library(factoextra)
```

```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_eig(pca, addlabels = TRUE)
```



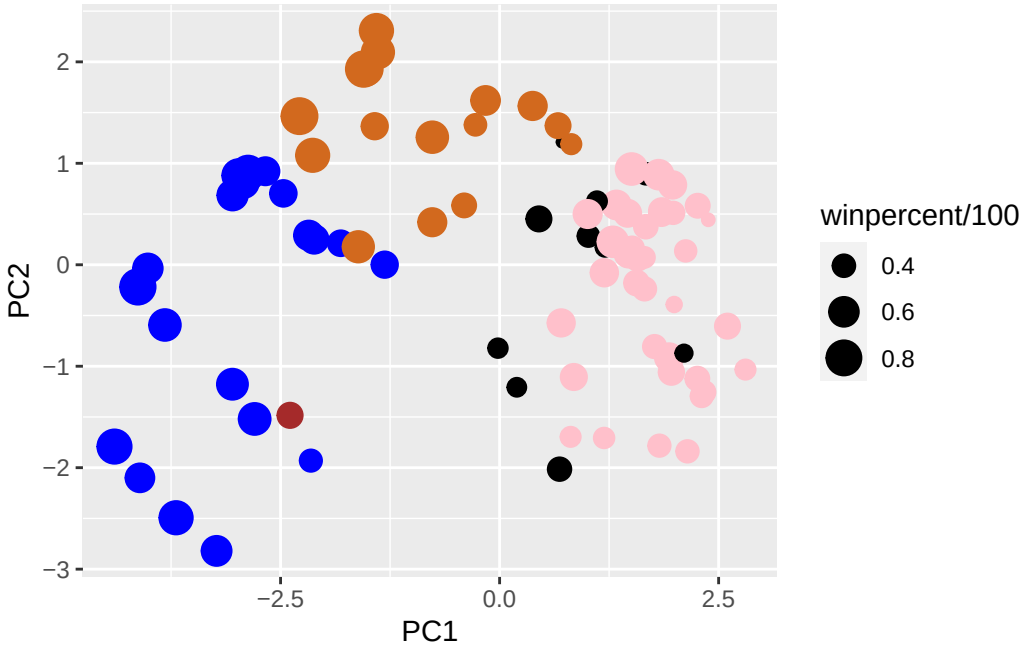

```
plot(pca$x[,1:2], col = my_cols, pch = 16)
```




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

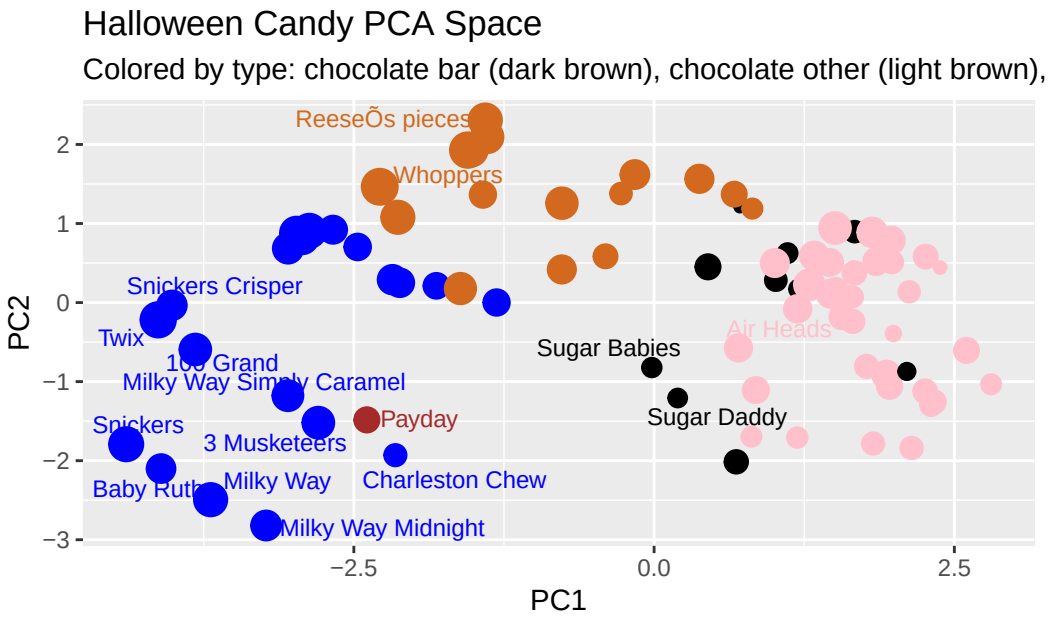
```
p <- ggplot(my_data) +
        aes(x = PC1, y = PC2,
            size = winpercent / 100,
            text = rownames(my_data),
            label = rownames(my_data)) +
        geom_point(col = my_cols)

p
```




```
p + geom_text_repel(size = 3.3, col = my_cols, max.overlaps = 5) +
  theme(legend.position = "none") +
  labs(title = "Halloween Candy PCA Space",
       subtitle = "Colored by type: chocolate bar (dark brown), chocolate other (light bro
       caption = "Data from 538")
```

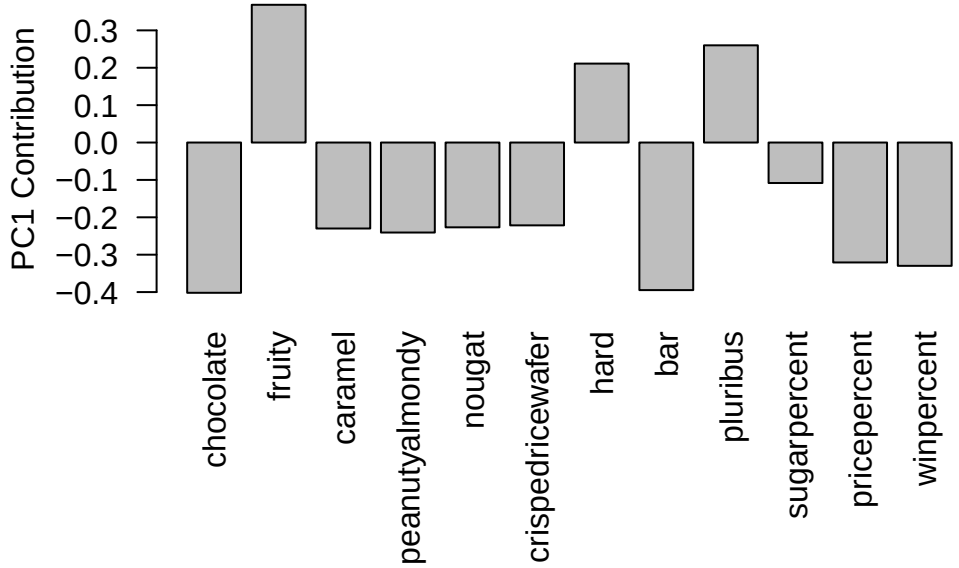
```
Warning: ggrepel: 69 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




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

# ggplotly(p)
```

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



Q24: Fruity, hard, and pluribus are in positive direction. Cutomarily it seems to make sense that fruity candies are hard candies packaged in bags or boxs.