class10

Muhammad Tariq

2025-05-01

##Setup

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)

# Load all necessary packages (assumes already installed)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(tinytex)
library(skimr)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(ggrepel)
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
```

###1. 1. Importing candy data

```
candy_file <- "candy-data.csv"
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
```

```
#Q1 How many different candy types are in the dataset?
nrow(candy)
```

```
## [1] 85
```

There are 85 different candy types in this dataset

```
# Q2: Number of fruity candy types
sum(candy$fruity == 1)
```

```
## [1] 38
```

There are 38 number of fruity candies

###2. What is your favorate candy?

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


candy["Air Heads", "winpercent"]
```

```
## [1] 52.34146
```

The winercent is 52% for Air Heads (fav candy)

```
#Q4. What is the winpercent value for “Kit Kat”?
candy["Kit Kat", "winpercent"]
```

```
## [1] 76.7686
```

The winpercent is 77% for kitkat

```
#Q5. What is the winpercent value for “Tootsie Roll Snack Bars”?
candy["Tootsie Roll Snack Bars", "winpercent"]
```

```
## [1] 49.6535
```

The winpercent is 50% for tootsie roll snack bars.

##Skim Function

```
library("skimr")
skim(candy)
```

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	▃▇▆▅▂

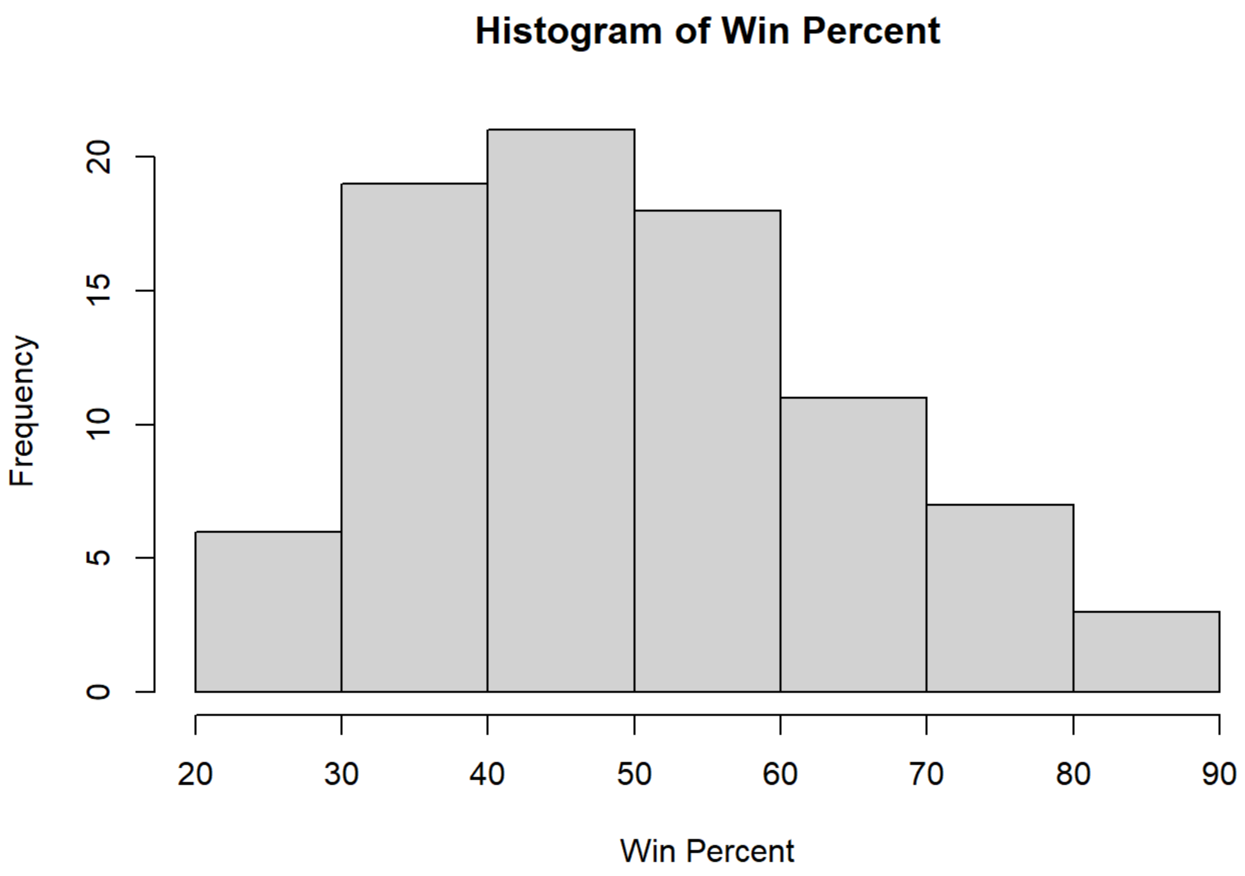
```
# Q6. Any variable on a different scale? (Check 'sugarpercent', 'pricepercent', 'winpercent')
# 'winpercent', 'sugarpercent', and 'pricepercent' are continuous, others are binary
# Confirms variable distributions

# Q7. 0 = does not have chocolate; 1 = has chocolate (binary)
# Q7. Interpretation of 0 and 1 in 'chocolate' column

# 0 = no chocolate, 1 = contains chocolate
table(candy$chocolate)  # Shows count of each
```

```
##
##  0  1
## 48 37
```

```
# Q8. Histogram of winpercent
hist(candy$winpercent, main = "Histogram of Win Percent", xlab = "Win Percent")
```



```
# Q9-Q10.
# Check symmetry visually
# Center (mean)
mean(candy$winpercent)
```

```
## [1] 50.31676
```

```
# Q11. Mean winpercent for chocolate vs fruity candies
mean(candy$winpercent[as.logical(candy$chocolate)])
```

```
## [1] 60.92153
```

```
mean(candy$winpercent[as.logical(candy$fruity)])
```

```
## [1] 44.11974
```

```
# Q12. Statistical test (Welch t-test)
t.test(candy$winpercent[as.logical(candy$chocolate)],
       candy$winpercent[as.logical(candy$fruity)])
```

```
##
##  Welch Two Sample t-test
##
## data:  candy$winpercent[as.logical(candy$chocolate)] and candy$winpercent[as.logical(candy$fr
uity)]
## 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
```

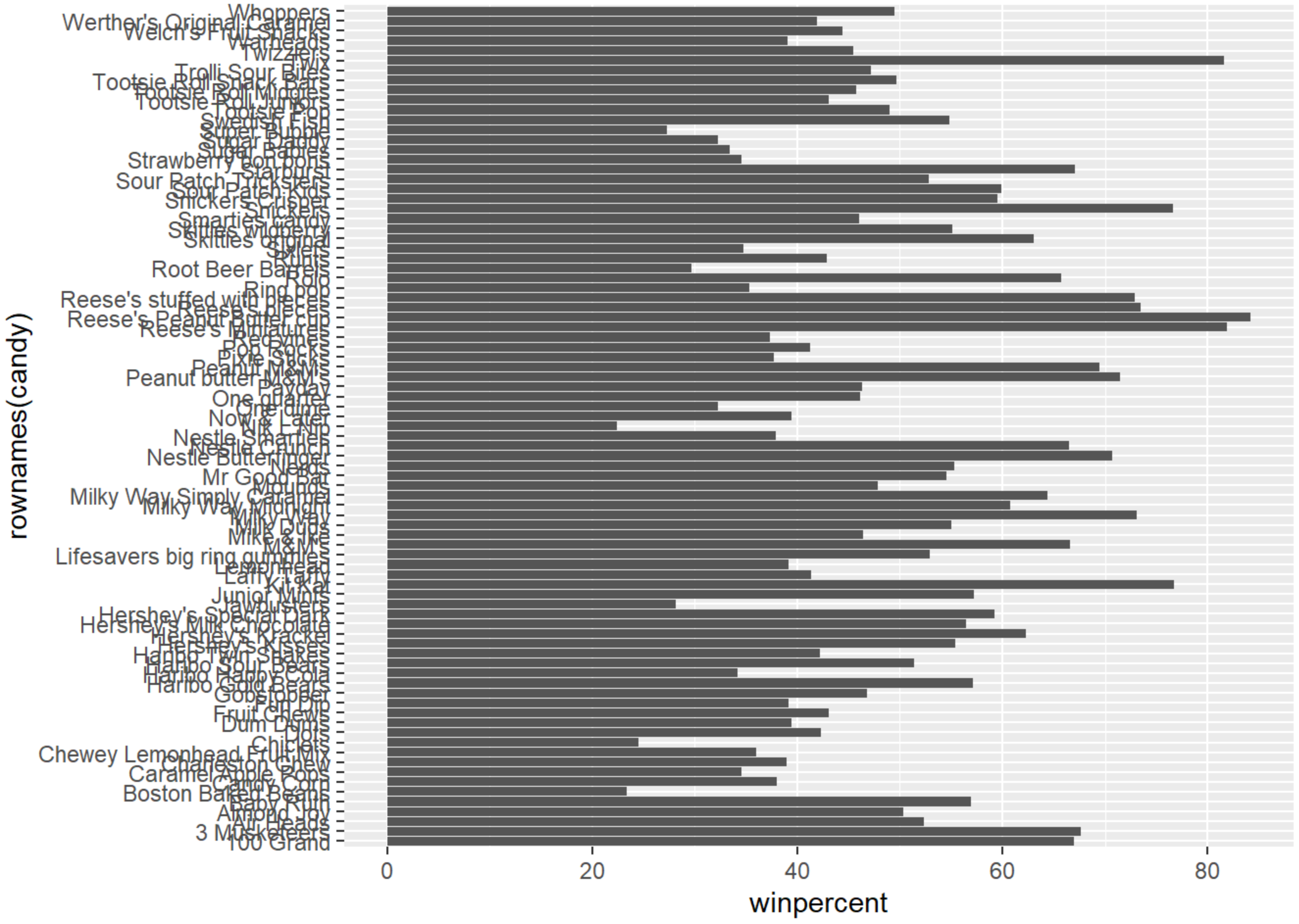
```
# Q13. 5 least liked candies
candy %>% arrange(winpercent) %>% head(5)
```

```
##                    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
```

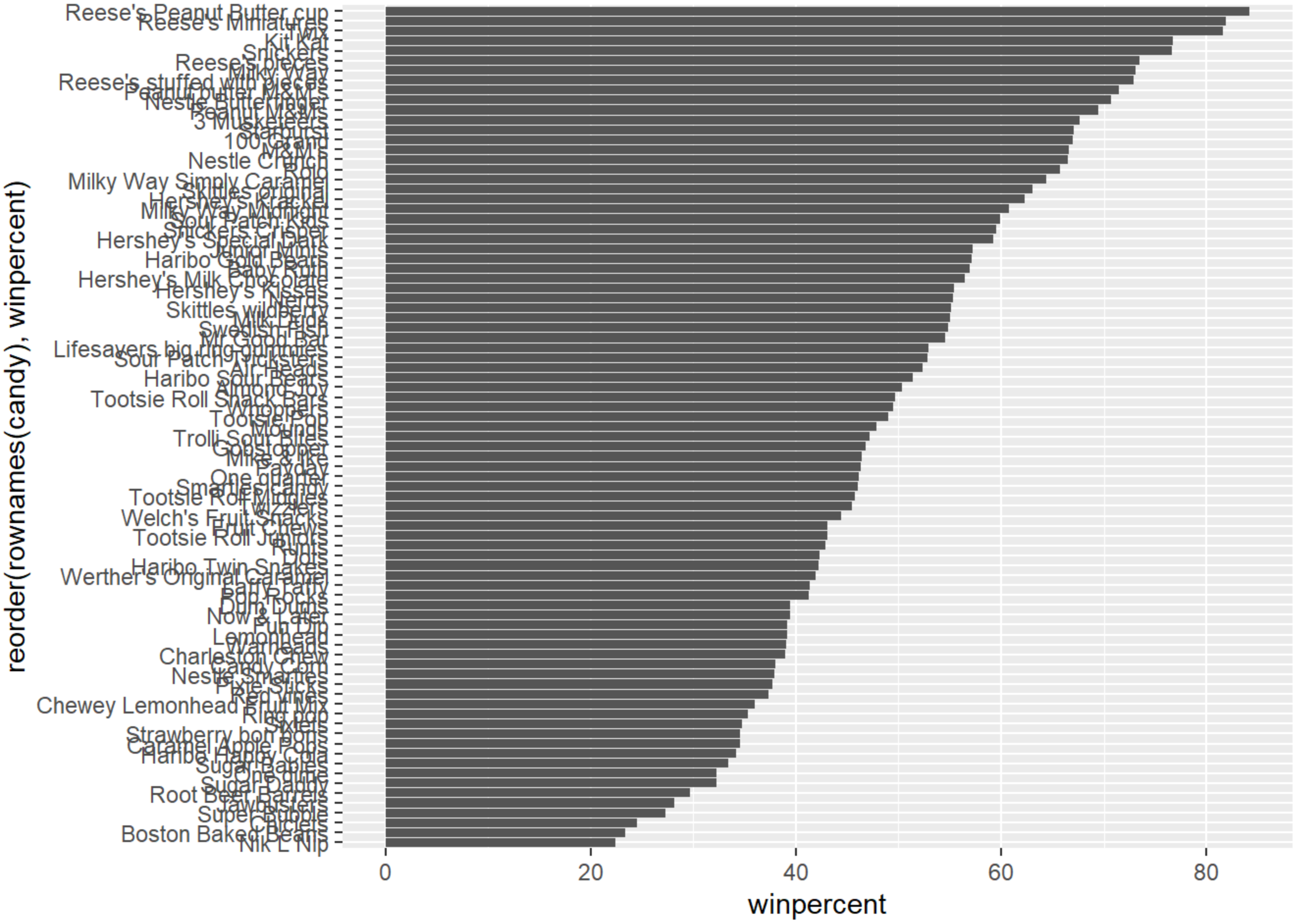
```
# Q14. 5 most liked candies
candy %>% arrange(desc(winpercent)) %>% head(5)
```

```
##                           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
```

```
# Q15. Barplot of winpercent
ggplot(candy) +
  aes(x = winpercent, y = rownames(candy)) +
  geom_col()
```

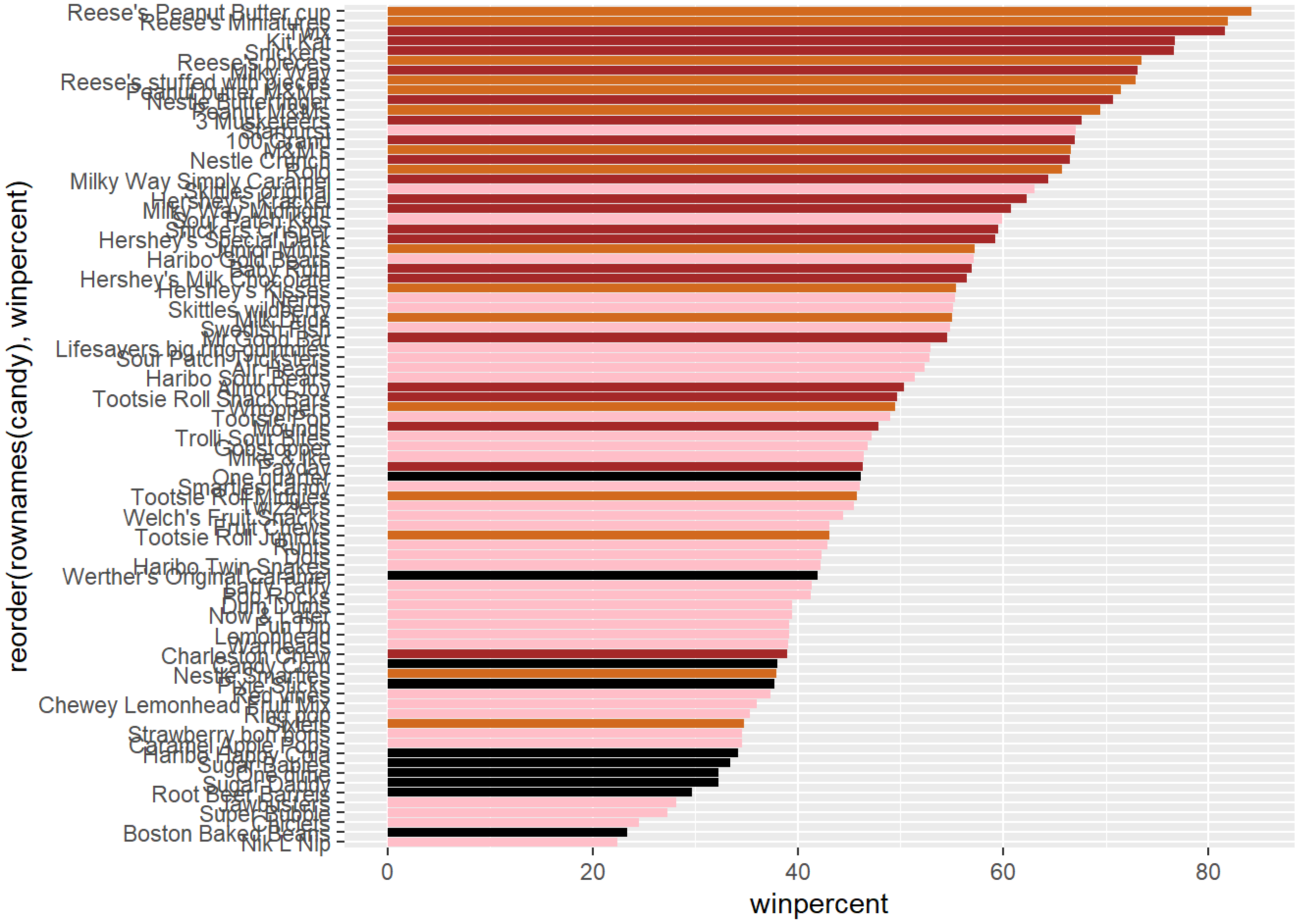


```
# Q16. Improve barplot with reorder
ggplot(candy) +
  aes(x = winpercent, y = reorder(rownames(candy), winpercent)) +
  geom_col()
```



```
# Color vector
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"

# Improved barplot with colors
ggplot(candy) +
  aes(x = winpercent, y = reorder(rownames(candy), winpercent)) +
  geom_col(fill = my_cols)
```



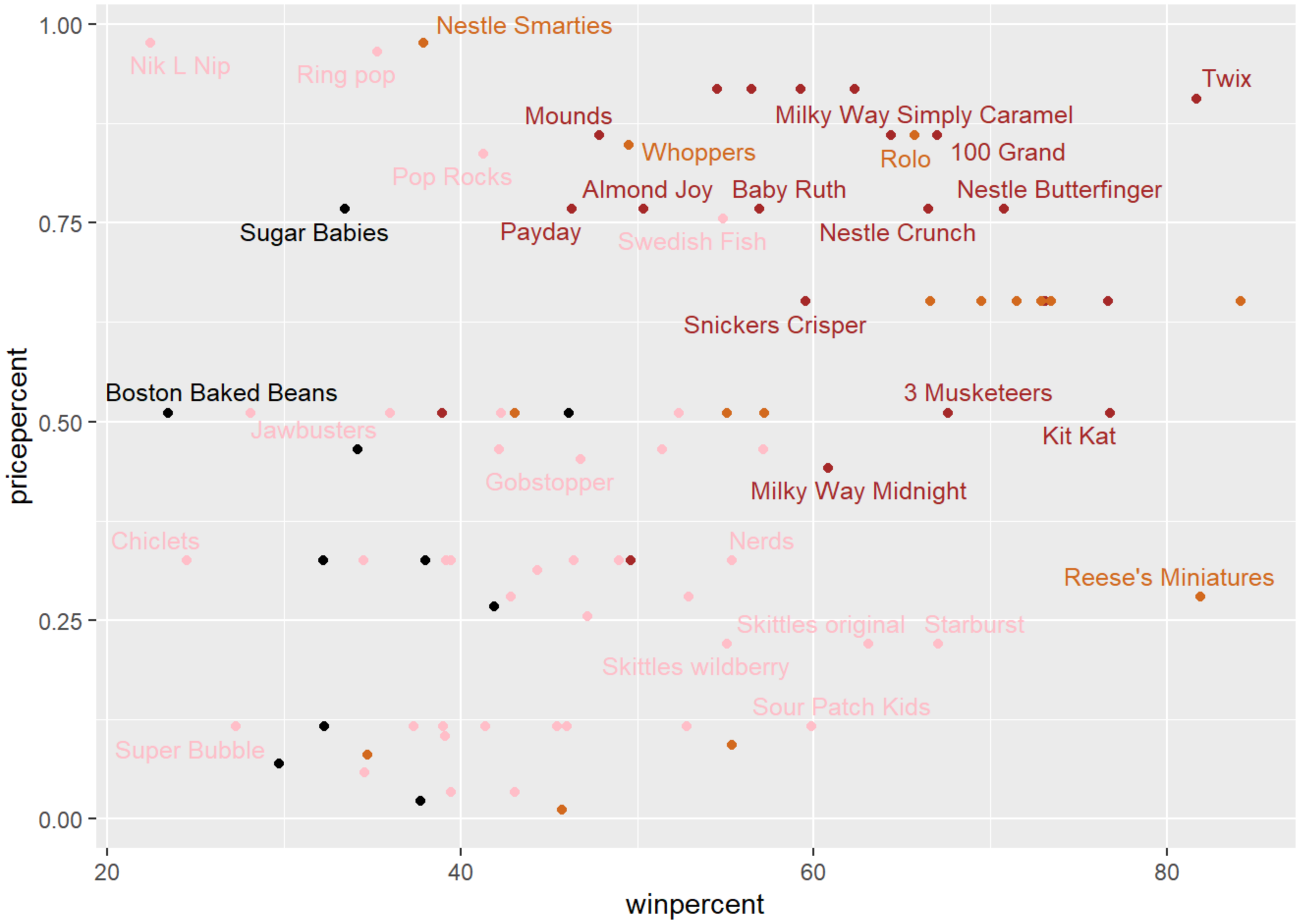
```
# Q17. Worst ranked chocolate candy
candy[as.logical(candy$chocolate), ] %>%
  arrange(winpercent) %>%
  head(1)
```

```
##         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. Best ranked fruity candy
candy[as.logical(candy$fruity), ] %>%
  arrange(desc(winpercent)) %>%
  head(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
```

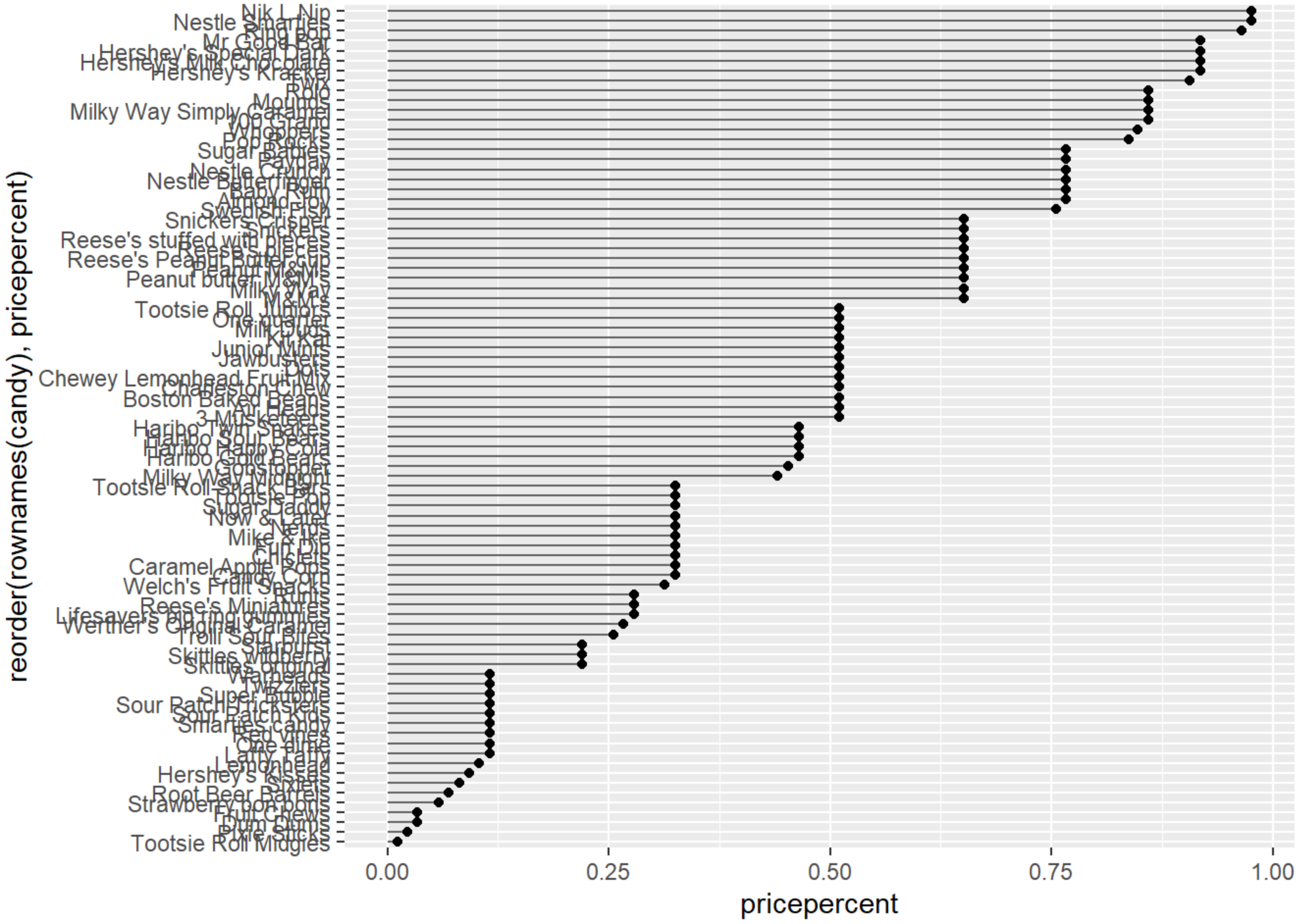
```
# Q19. Best value candy (high winpercent, low pricepercent)
ggplot(candy) +
  aes(x = winpercent, y = pricepercent, label = rownames(candy)) +
  geom_point(col = my_cols) +
  geom_text_repel(col = my_cols, size = 3.3, max.overlaps = 5)
```



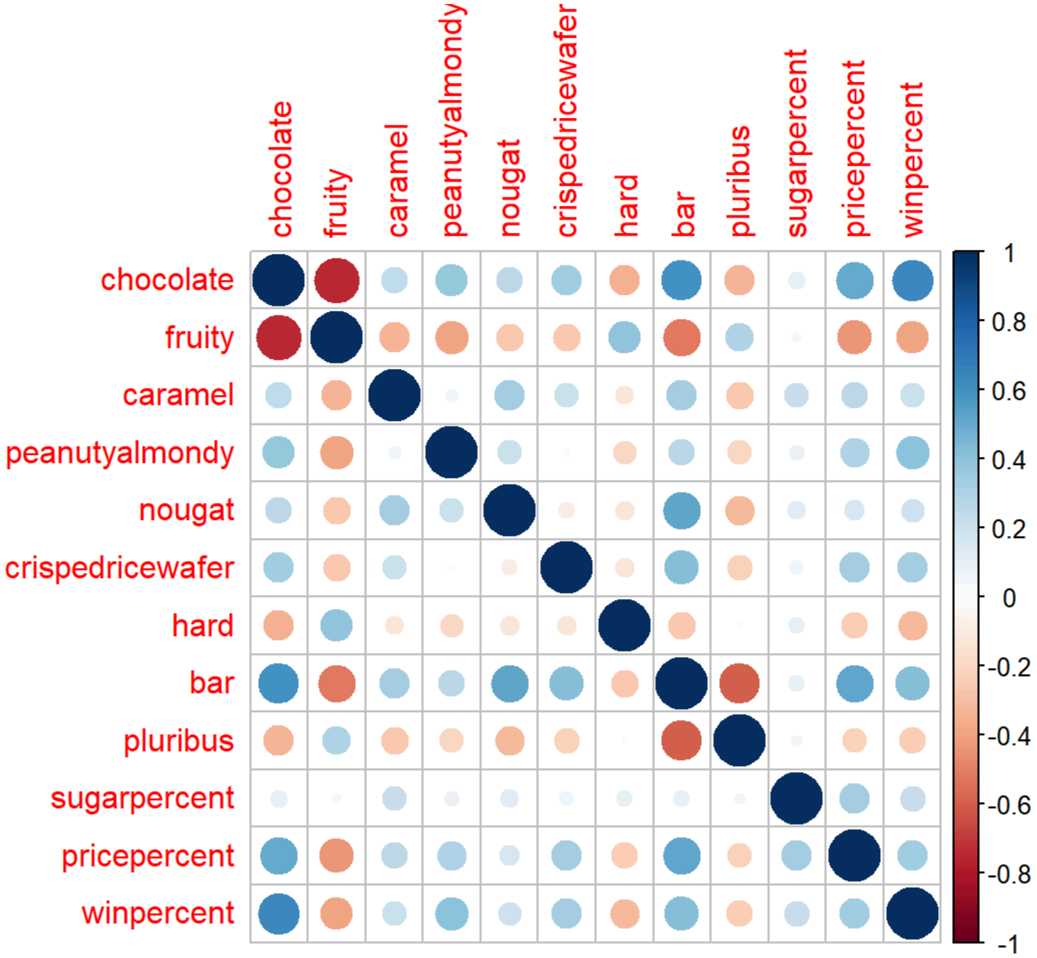
```
# Q20. Top 5 most expensive candies + least popular among them
ord <- order(candy$pricepercent, decreasing = TRUE)
head(candy[ord, c("pricepercent", "winpercent")], n = 5)
```

```
##                           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
```

```
# Q21. Lollipop chart
ggplot(candy) +
  aes(x = pricepercent, y = reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(xend = 0, yend = reorder(rownames(candy), pricepercent)),
               col = "gray40") +
  geom_point()
```



```r
# Q22-Q23. Correlation matrix
cij <- cor(candy)
corrplot(cij)
```

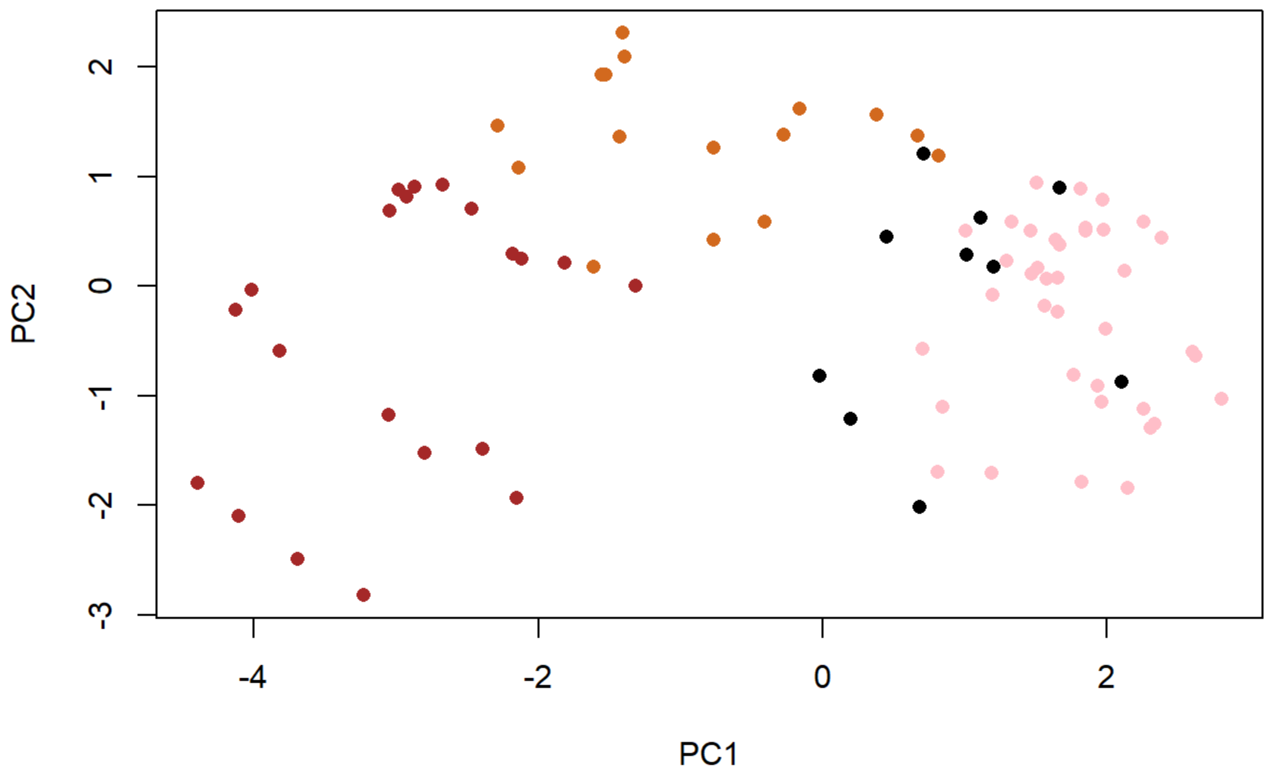


```
#The most positively correlated variables are chocolate and bar. This makes sense because chocol
ate candies are often in bar form. You can verify this in the corrplot by looking at the darkes
t/highest value.
```

```
# Q24. 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
```

```
# Base R plot of PC1 vs PC2
plot(pca$x[, 1:2], col = my_cols, pch = 16)
```

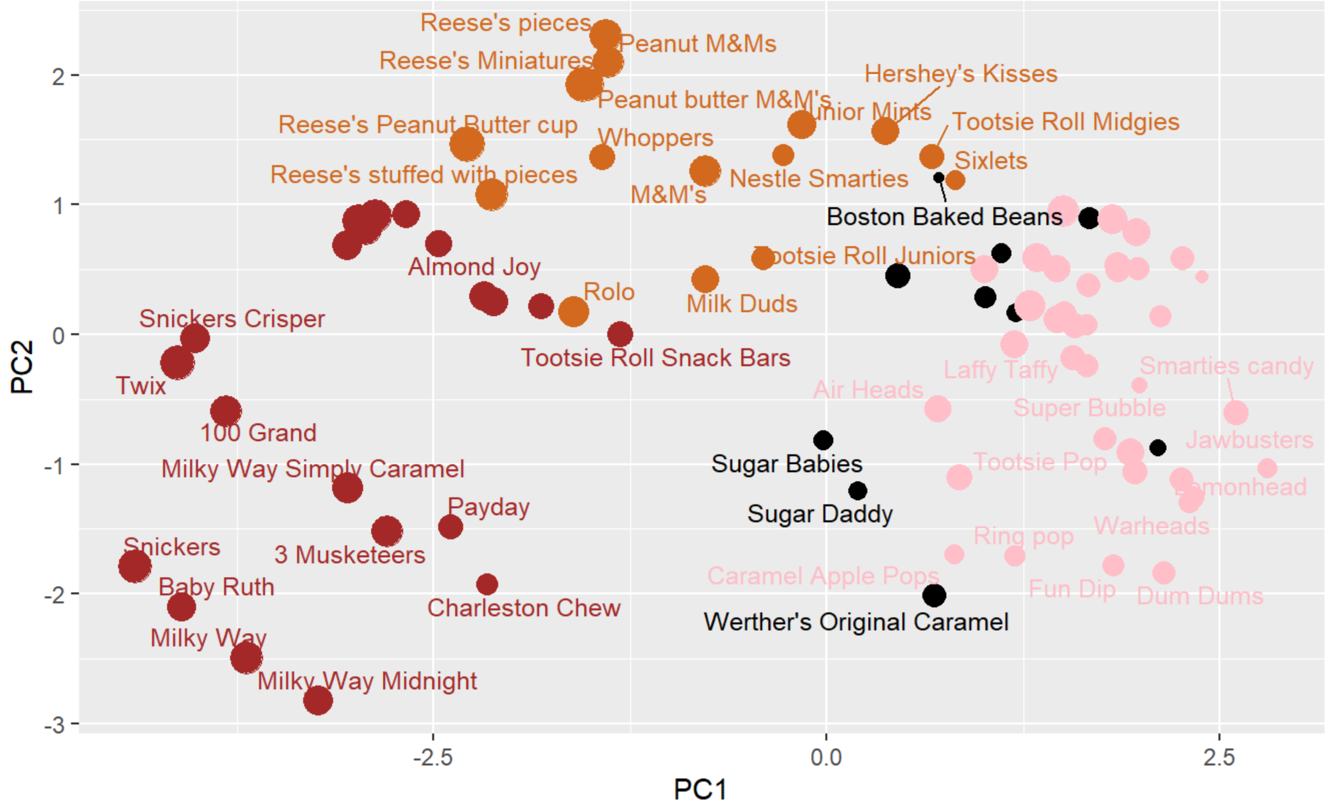


```
# ggplot PCA
my_data <- cbind(candy, pca$x[, 1:3])
p <- ggplot(my_data) +
  aes(x = PC1, y = PC2,
      size = winpercent / 100,
      text = rownames(my_data),
      label = rownames(my_data)) +
  geom_point(col = my_cols)

# PCA plot with 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), frui
ty (red), other (black)",
    caption = "Data from 538"
  )
```

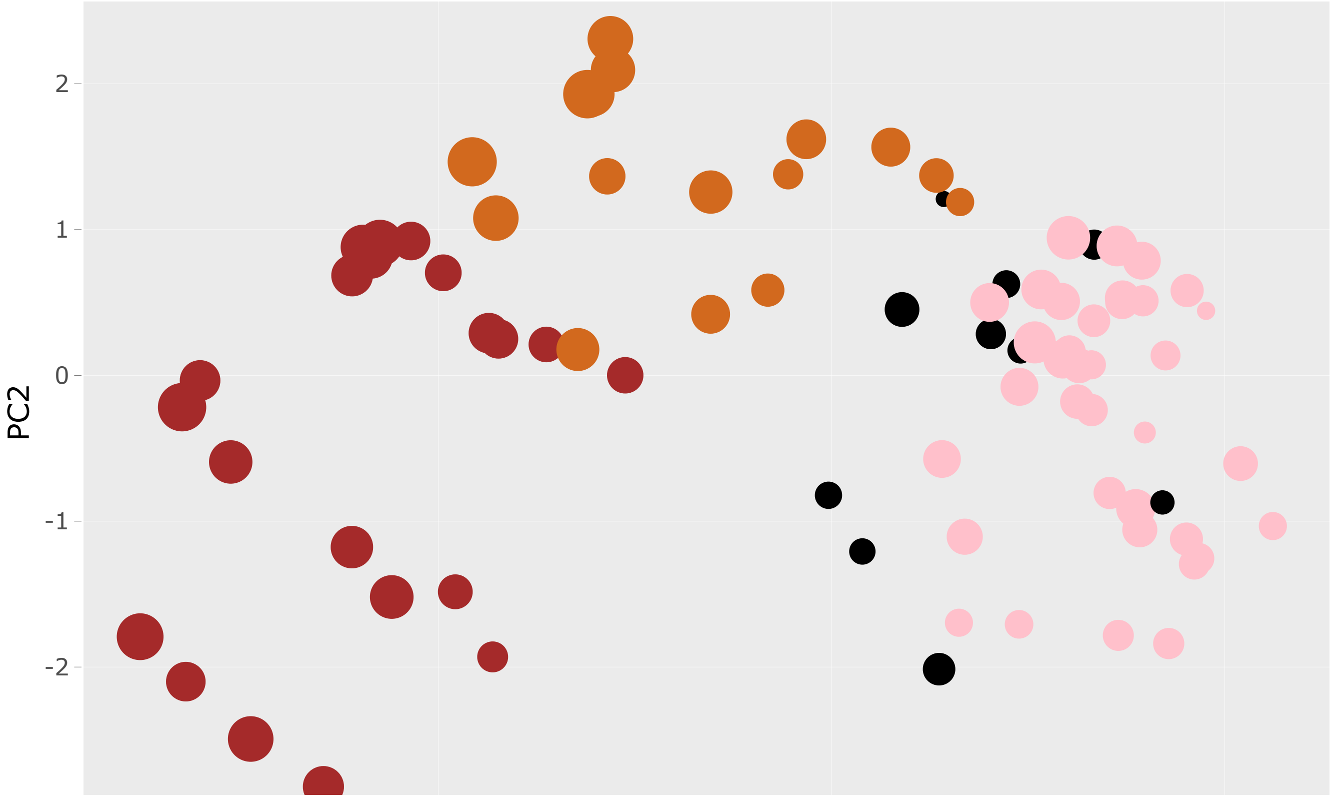
Halloween Candy PCA Space

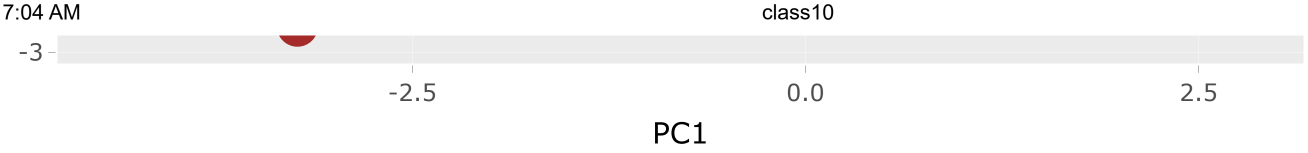
Colored by type: chocolate bar (dark brown), chocolate other (light brown), fruity (red), other (black



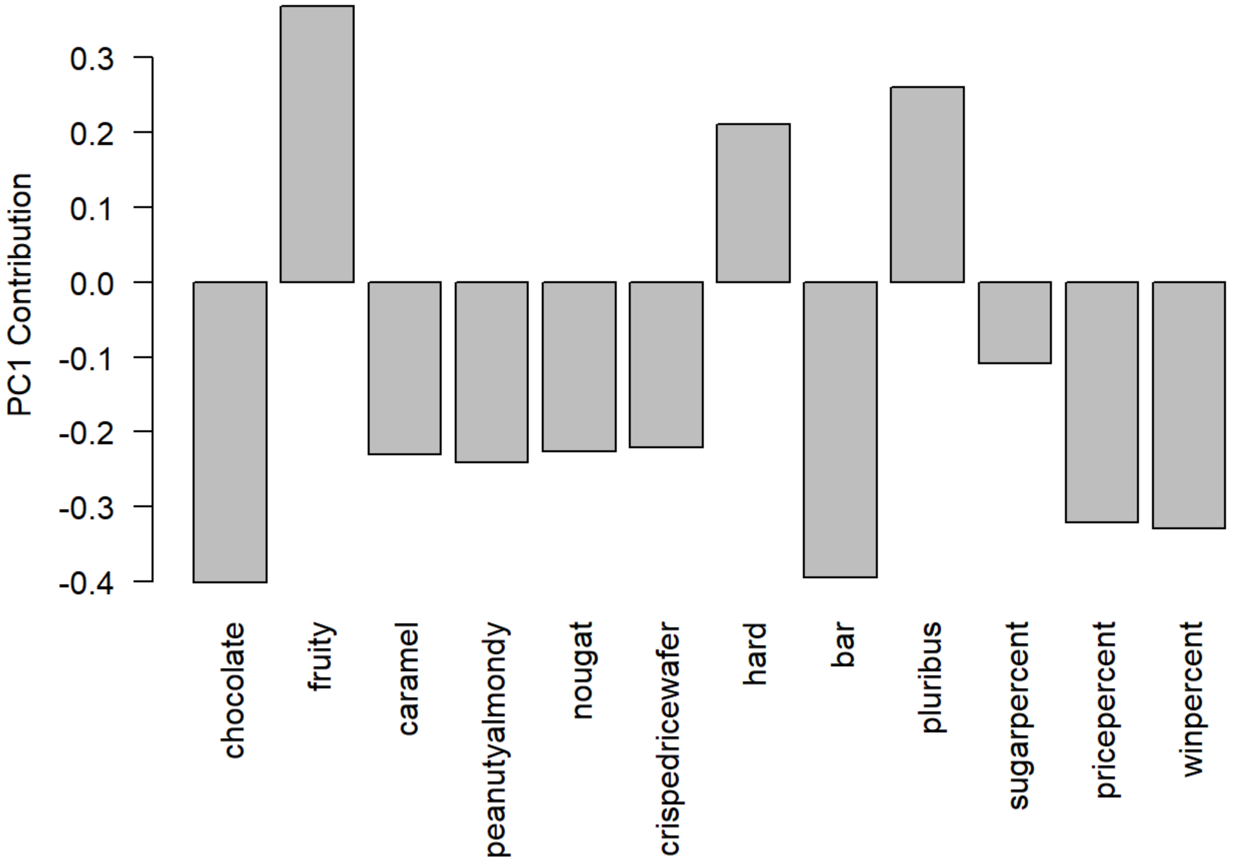
Data from 538

```
# Optional: interactive plot
ggplotly(p)
```





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


PC1 Contribution
chocolate
fruity
caramel
peanutyalmondy
nougat
crispedricewafer
hard
bar
pluribus
sugarpercent
pricepercent
winpercent