halloween

2023-05-15

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

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
```

Q1. How many different candy types?: 85 different candies, 9 different types?

Q2. How many fruity candy types are in the dataset? 38 fruity type candies

```
sum(candy[, "fruity"])
```

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

```
## Winpercent of Twix
candy["Twix", ]$winpercent
```

```
## [1] 81.64291
```

Q3. Favorite candy and winpercent? It's Kit Kat.

Q4. Winpercent for Kit Kat?

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

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

Q5. Winpercent for Tootsie Rolls?

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

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

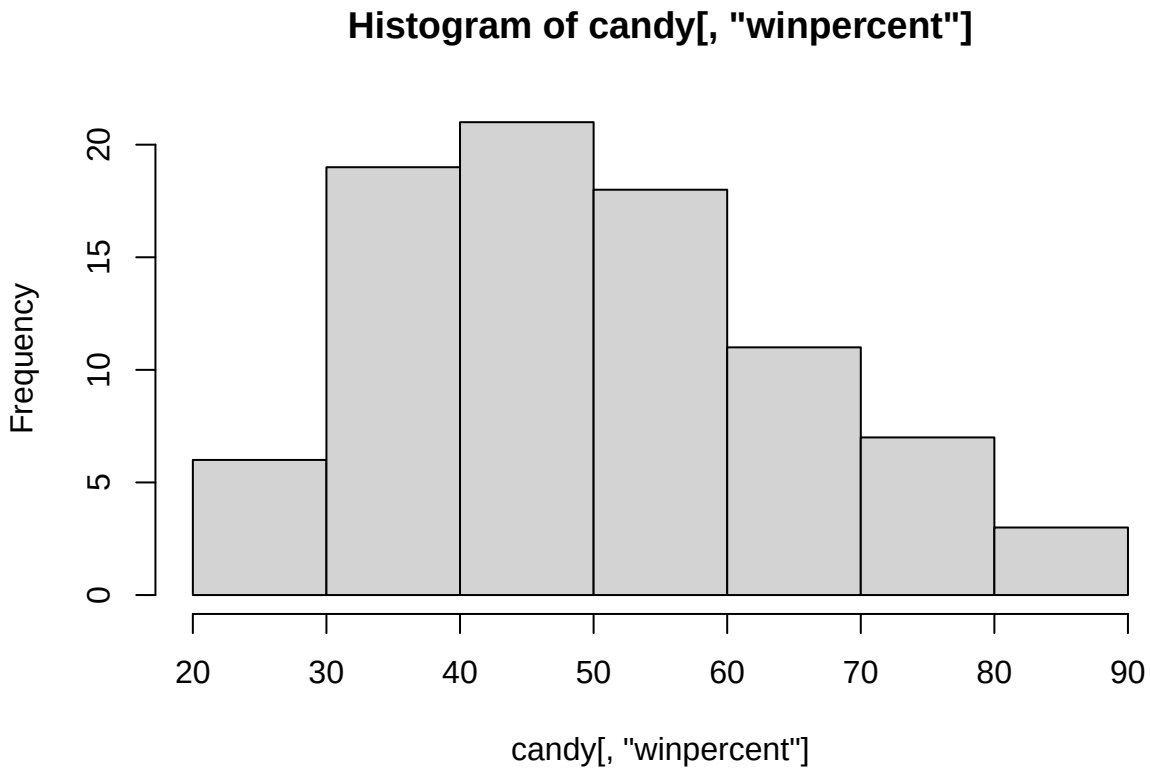
```
library("skimr")
#skim(candy)
```

Q6. Any variable/column that is on a different scale to the other data? While most of the variables are in a scale of 0 to 1, "winpercent" is the only variable with a scale from 0 to 100.

Q7. What does the O and 1 represent in the chocolate column? 0 means the candy does not contain chocolate. 1 means the candy contains chocolate.

Q8. Histogram of winpercent

```
hist(candy[, "winpercent"])
```



Q9. Is the distribution symmetrical? No. It is slightly skewed to the right.

Q10. Is the center of distribution above or below 50? It is below 50 as it is right skewed.

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

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

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

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

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

Chocolate is ranked higher on average.

Q12. Is this difference statistically significant?

With a p-value below 0.05, this is statistically significant.

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

```
tail(candy[order(candy$winpercent),], n=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
## Reese's Miniatures                 1      0       0              1      0
## Reese's 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
## Reese's Miniatures                         0    0   0        0        0.034
## Reese's 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
## Reese's Miniatures                0.279   81.86626
## Reese's Peanut Butter cup         0.651   84.18029
```

```
head(candy[order(candy$winpercent),], n=5)
```

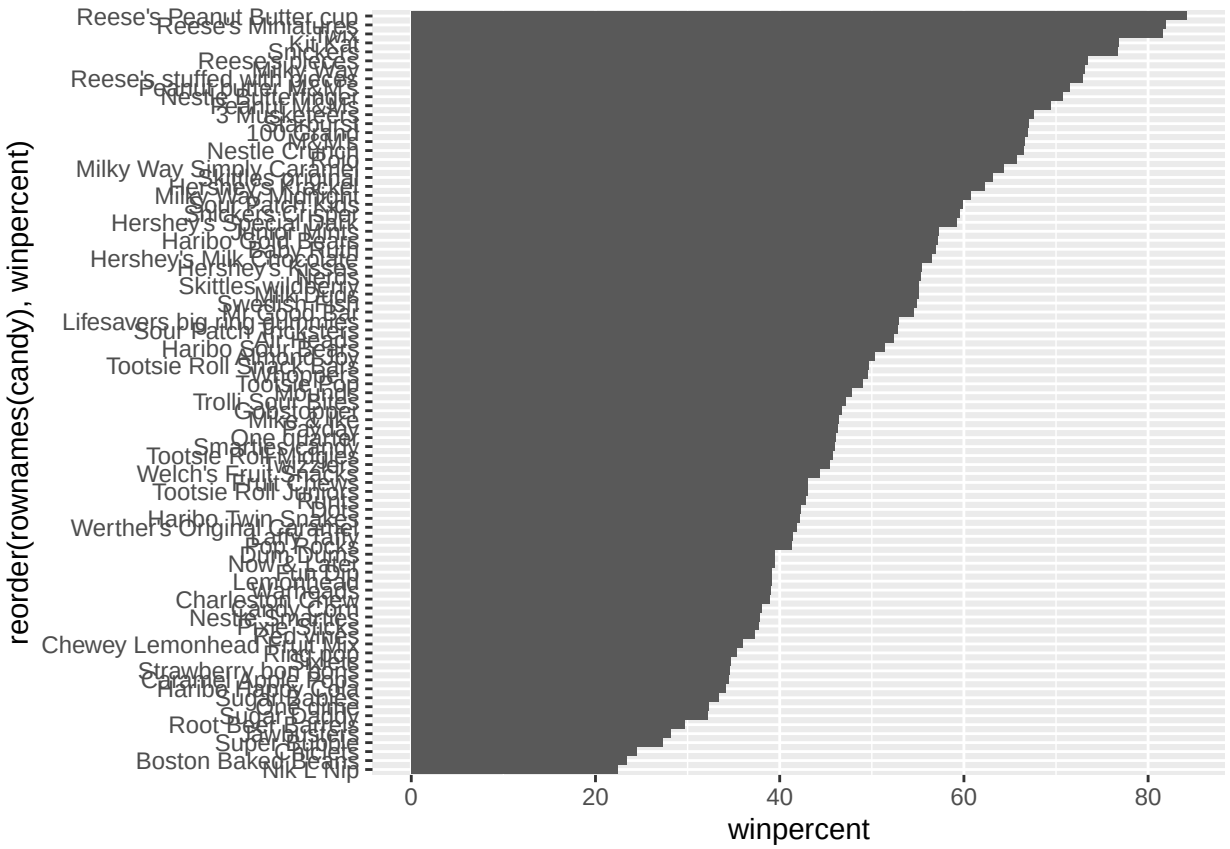
```
##                    chocolate fruity caramel peanutyalmondy nougat
## Nik L Nip                  0      1       0              0      0
```

```
## 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
```

Q13. Five least liked candy types? Nik L Nip, Boston Baked Beans, Chiclets, Supper Bubble, Jawbusters

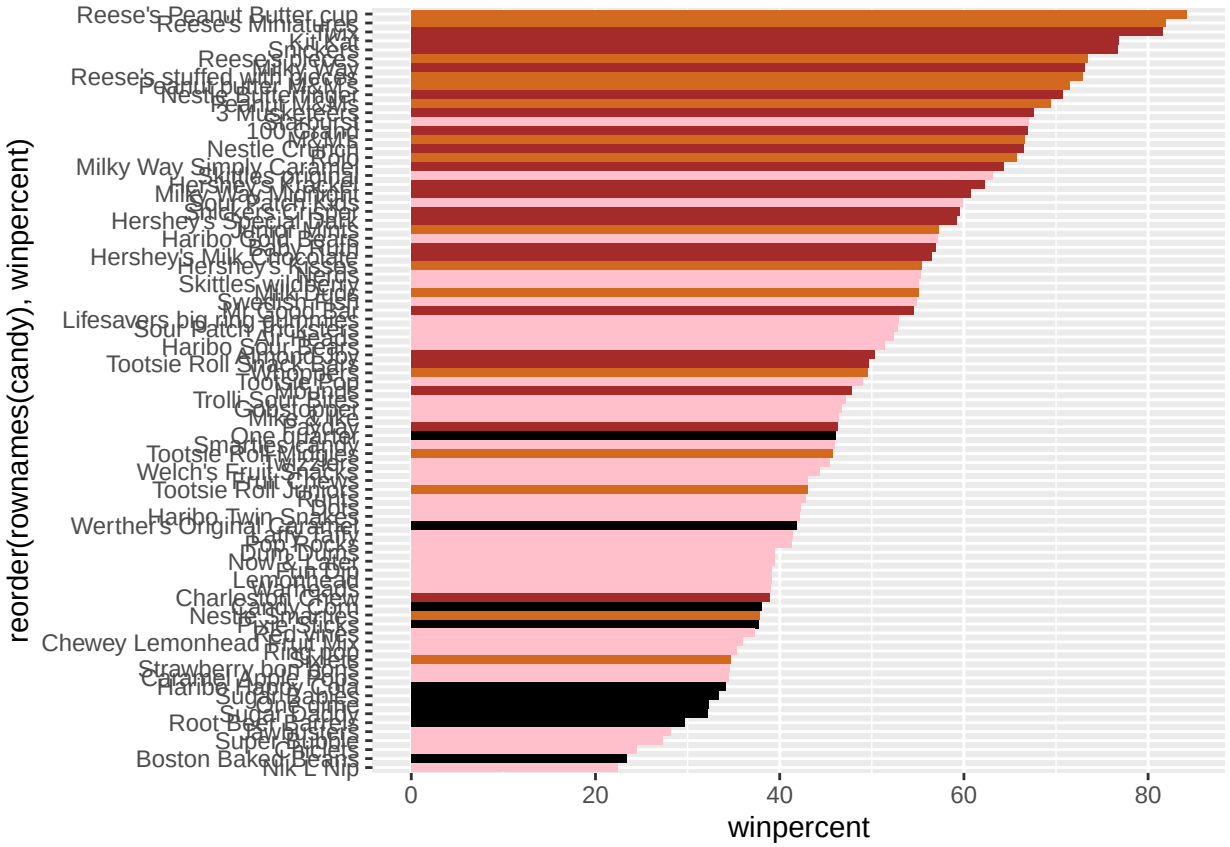
Q14. Top five candies? Snickers, Kit Kat, Twix, Reese's Mini, Reese's Peanut Butter Cups

```
library(ggplot2)

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



```
##setting up color vectors for the plot
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"
```

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

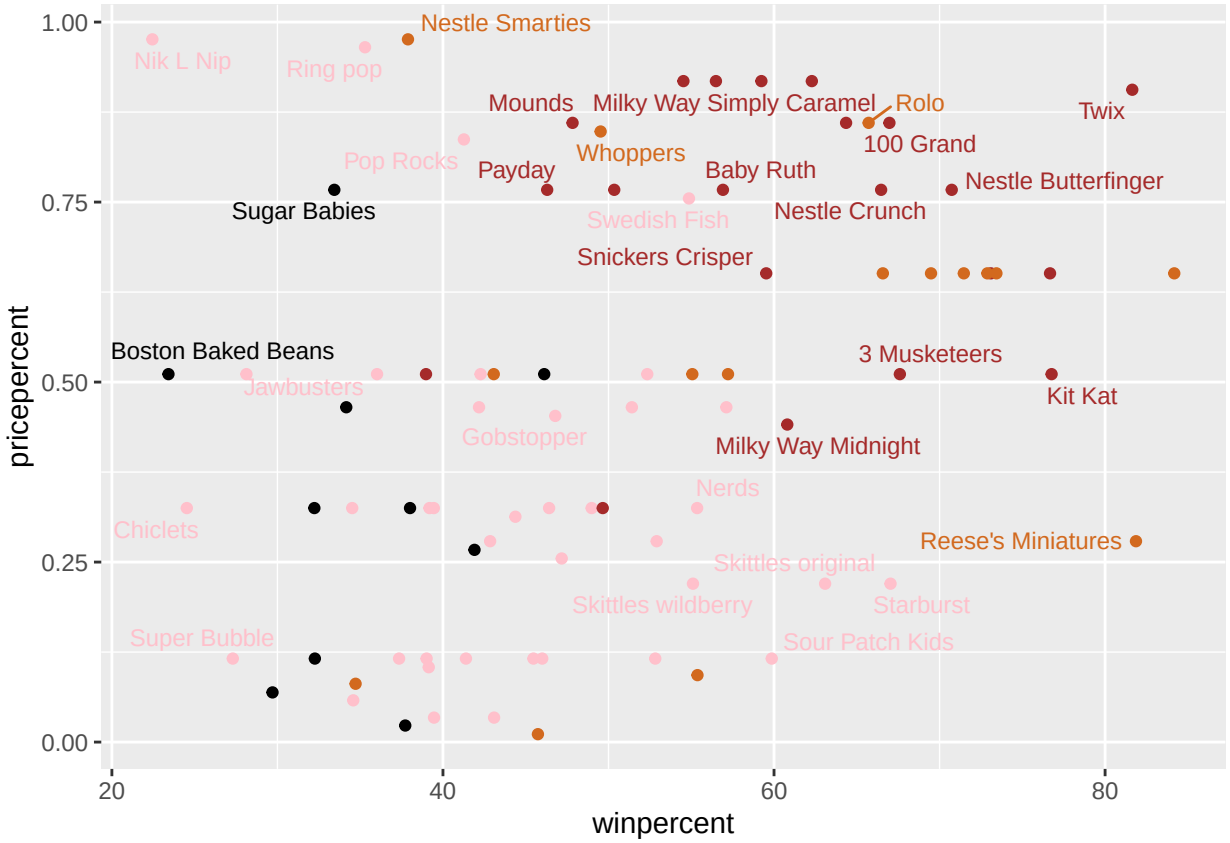


Q17. Worst ranked chocolate candy? Sixlets

Q18. Best ranked fruity candy? Starburst

```
library(ggrepel)

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




```
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
```

```
tail(candy[ord, c(11, 12)], n=5)
```

```
##                        pricepercent winpercent
## Strawberry bon bons           0.058   34.57899
## Dum Dums                      0.034   39.46056
## Fruit Chews                   0.034   43.08892
## Pixie Sticks                  0.023   37.72234
## Tootsie Roll Midgies          0.011   45.73675
```

Q19. Which candy is highest rank in terms of winpercent for the least money?

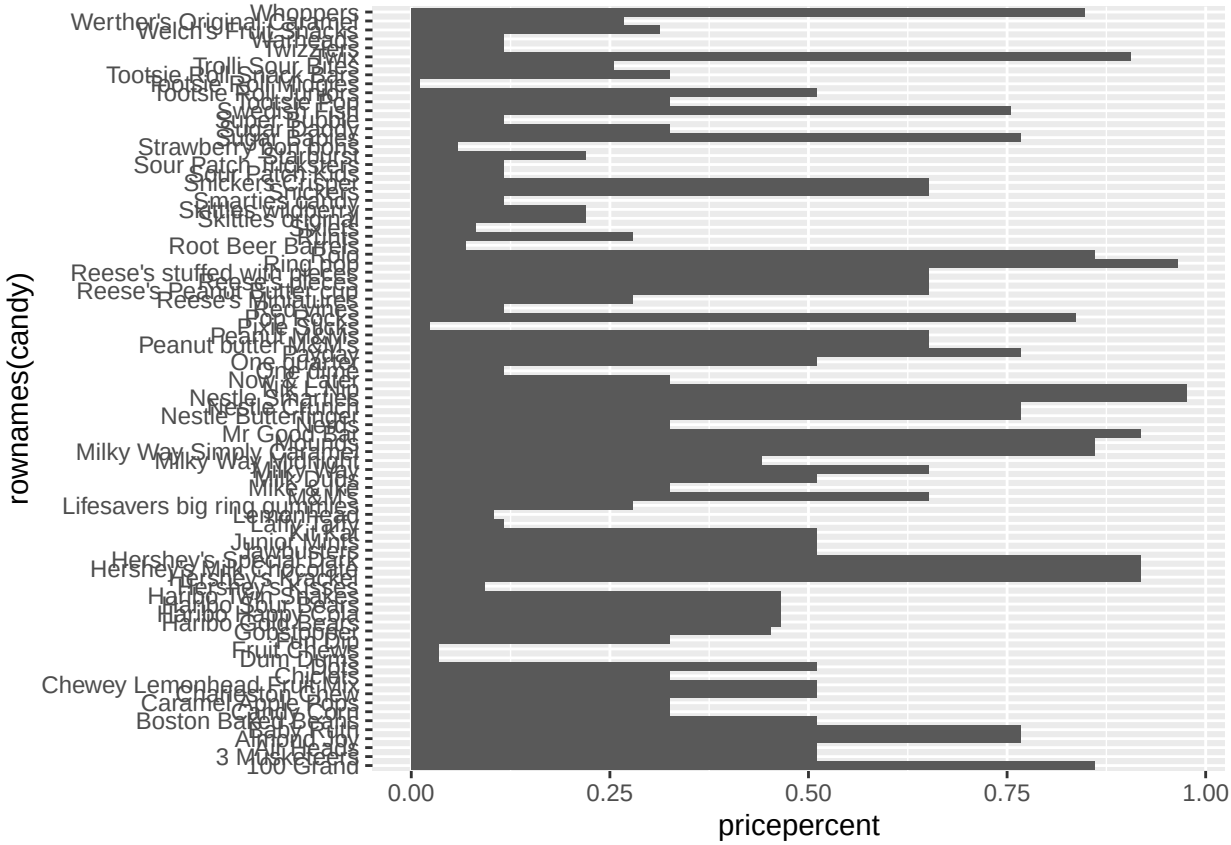
Reese's Miniatures are the bang for your buck.

Q20. Top 5 most expensive candies and which is least popular?

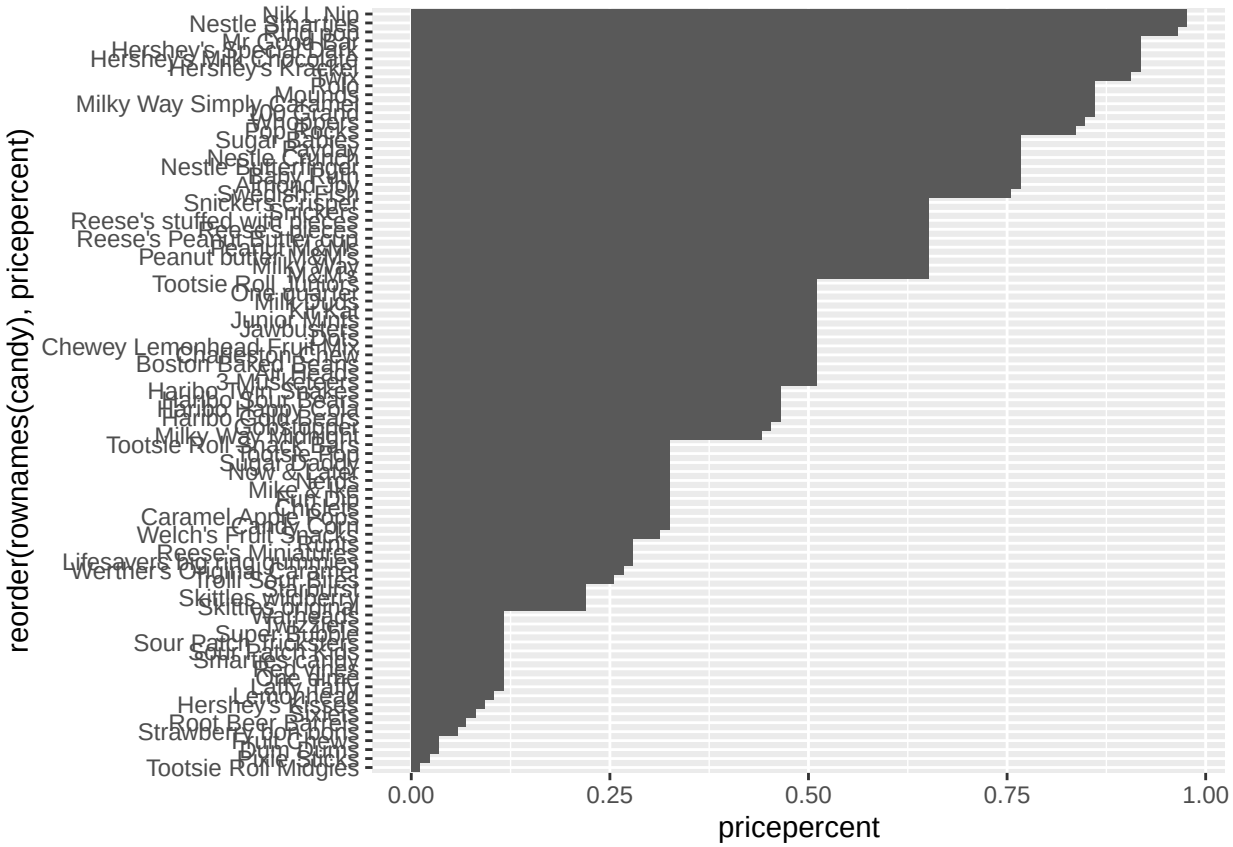
Nik L Nip, Nestle Smarties, Ring Pop, Hershey's Krackel, Hershey's Milk Choc

Nik L Nip are the least popular.

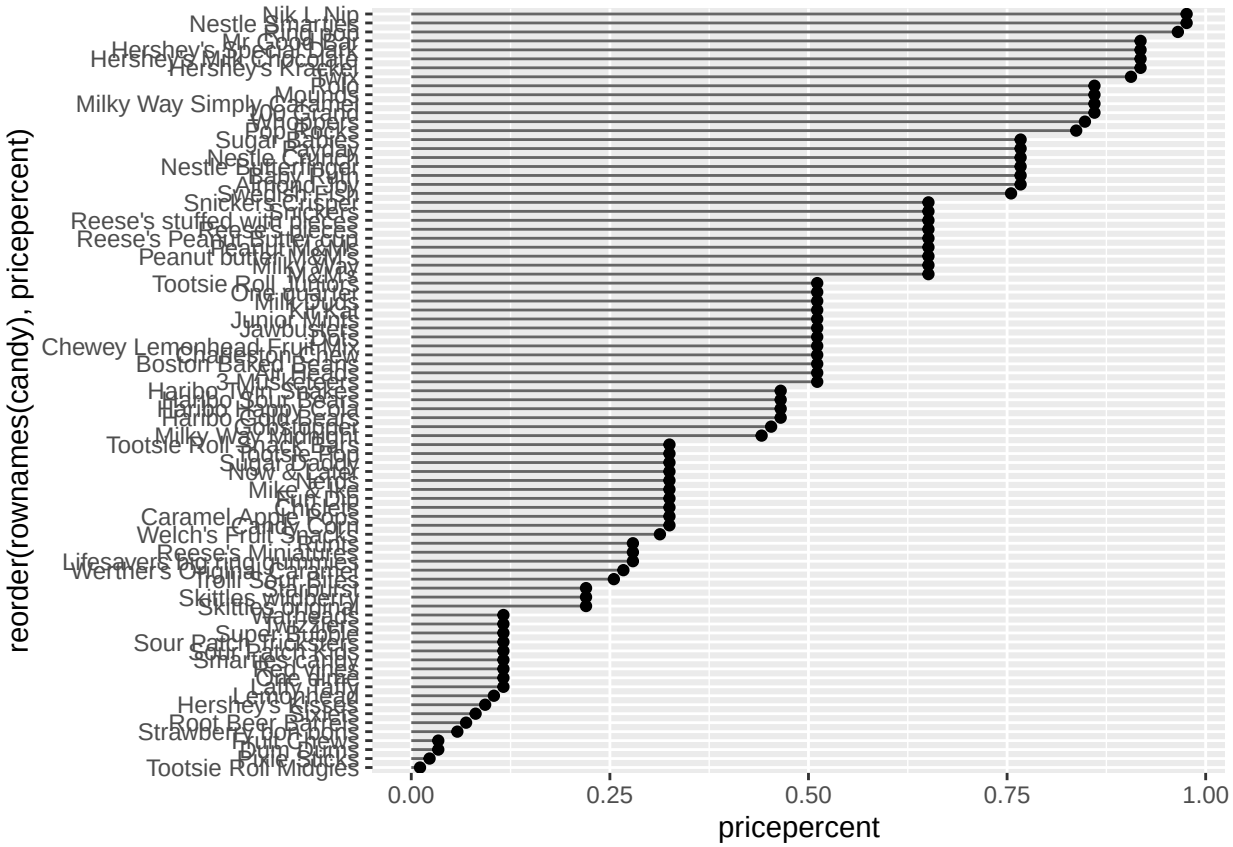
```
##first plot
ggplot(candy) +
  aes(pricepercent, rownames(candy)) +
  geom_col()
```




```
#reordering in descending order
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy),pricepercent)) +
  geom_col()
```



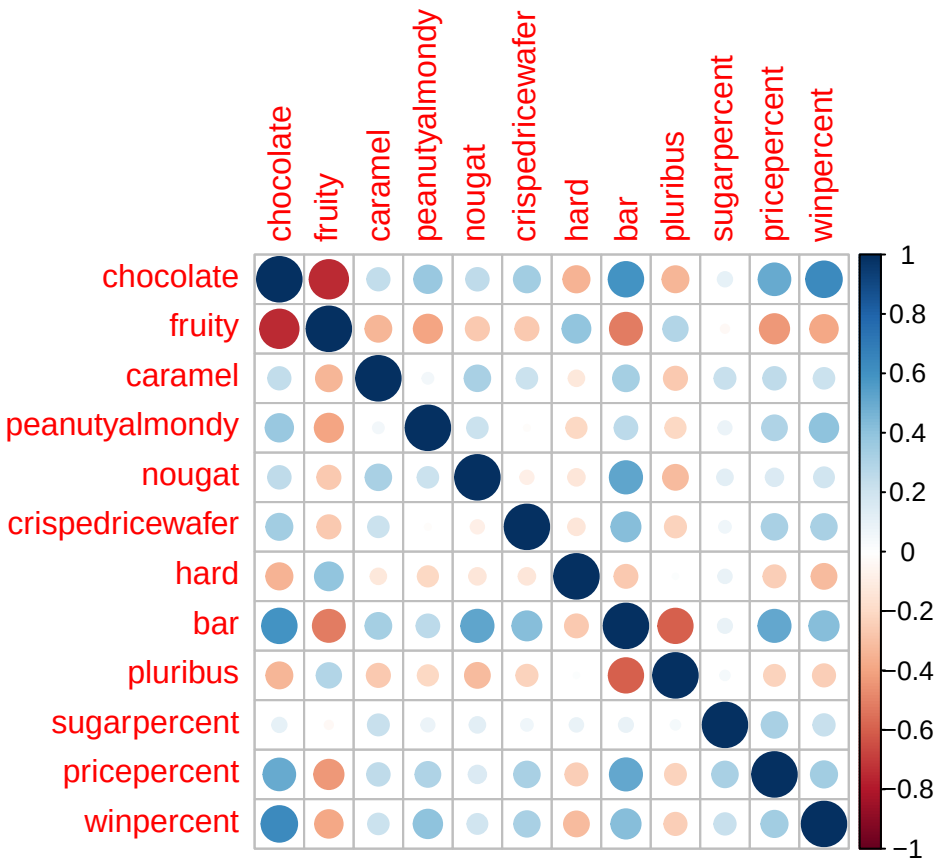

```
##col to segment + dot lollipop plot
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
  geom_point()
```




```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

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



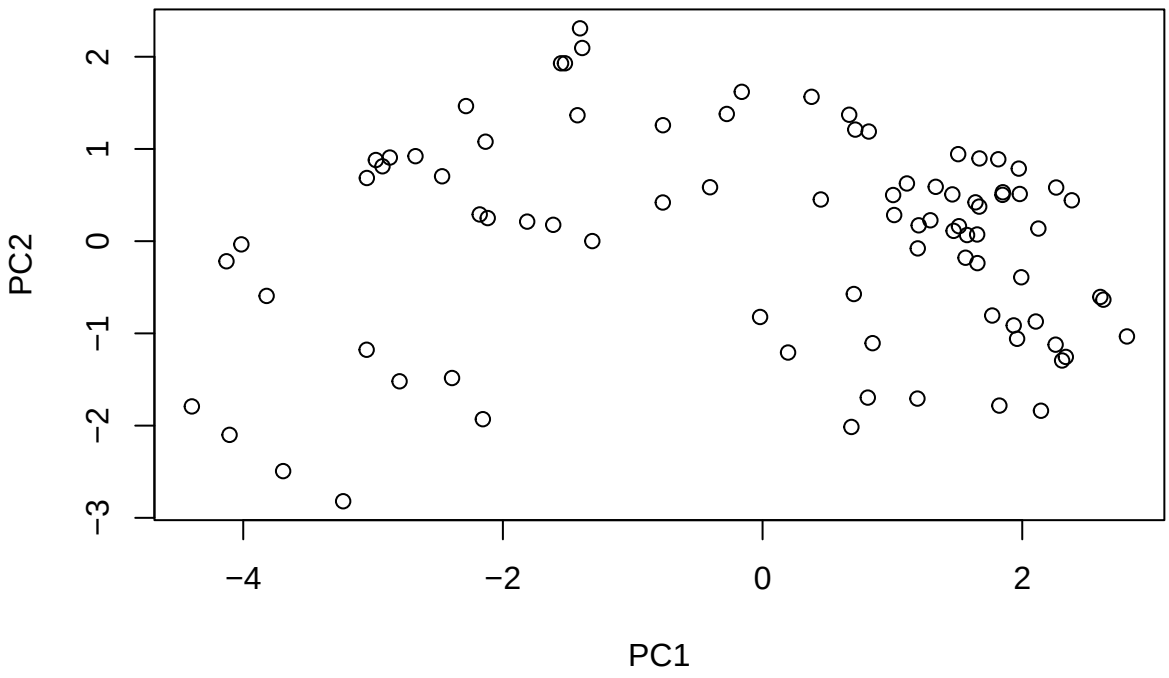
Q22. What two variables are anti-correlated? Chocolate and Fruit [Weird, I love chocolate + raspberry]

Q23. Which two variables are most positively correlated? Chocolate and Bar. Great form for chocolate to come in.

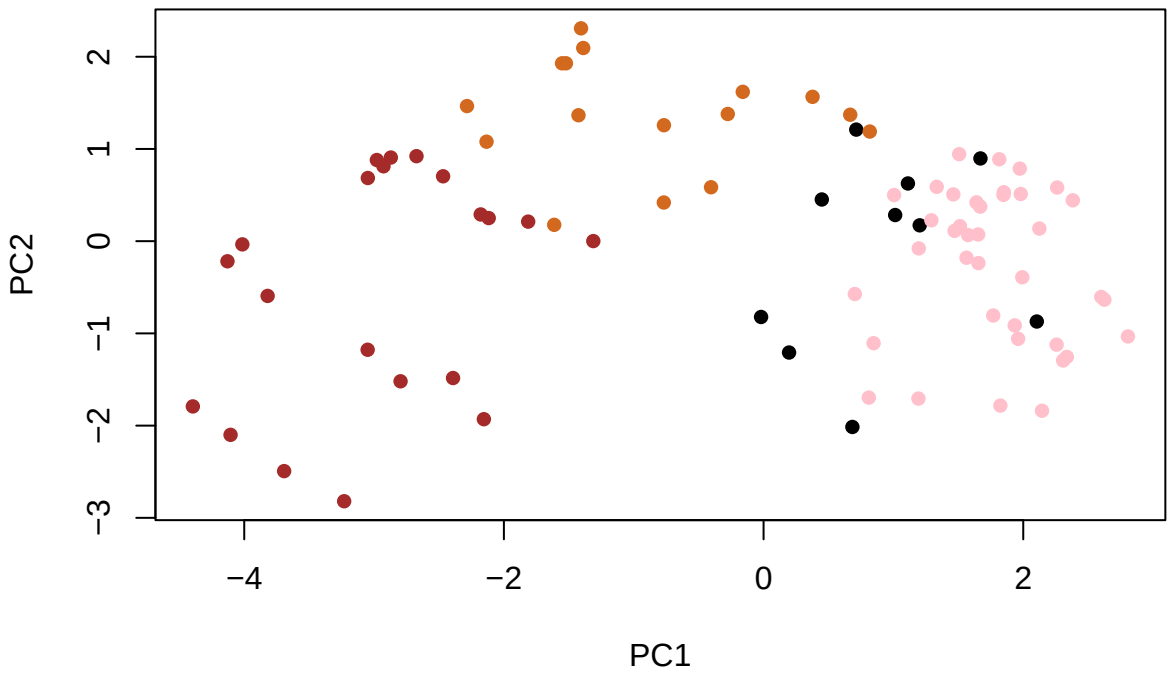
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
## 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
```

```
##plotting pca score plot pc1 v pc2
plot(pca$x[,1:2])
```

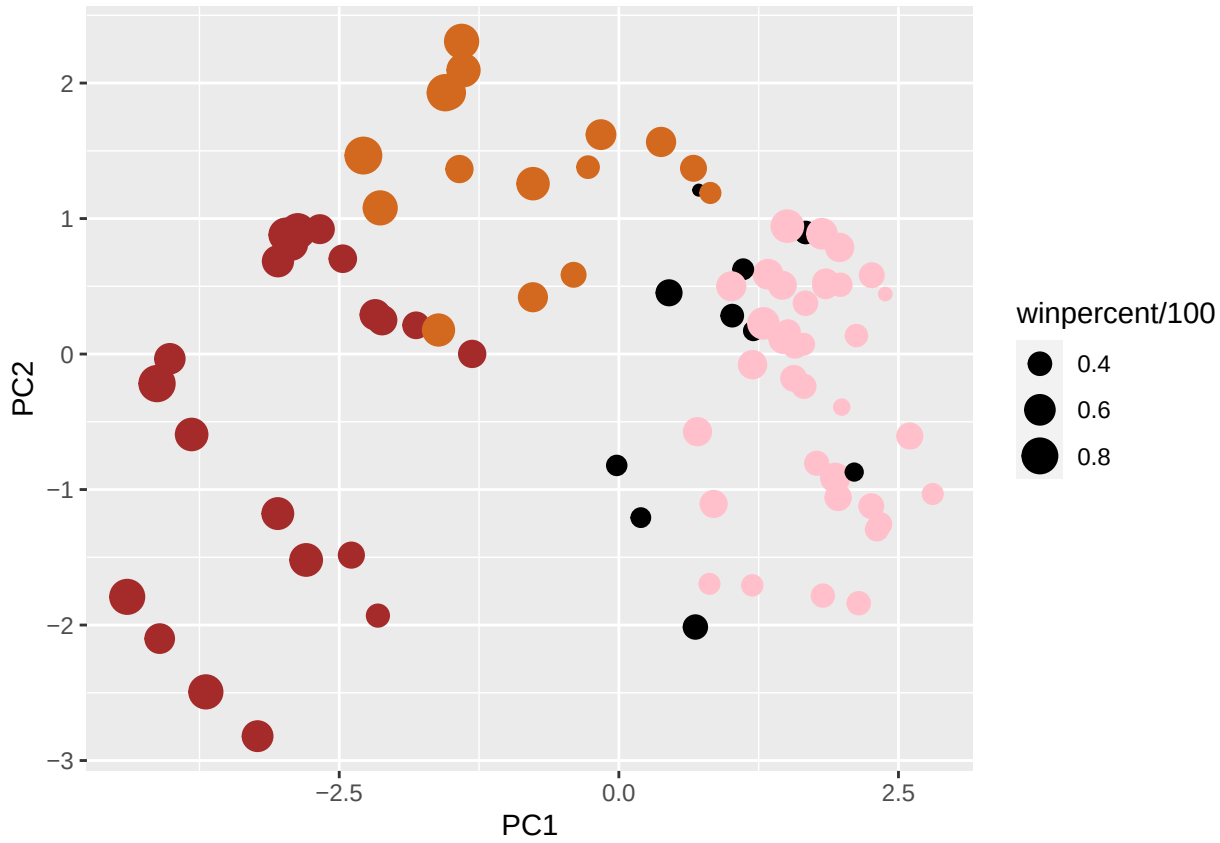



```
##adding color to plot
plot(pca$x[,1:2], col=my_cols, pch=16)
```



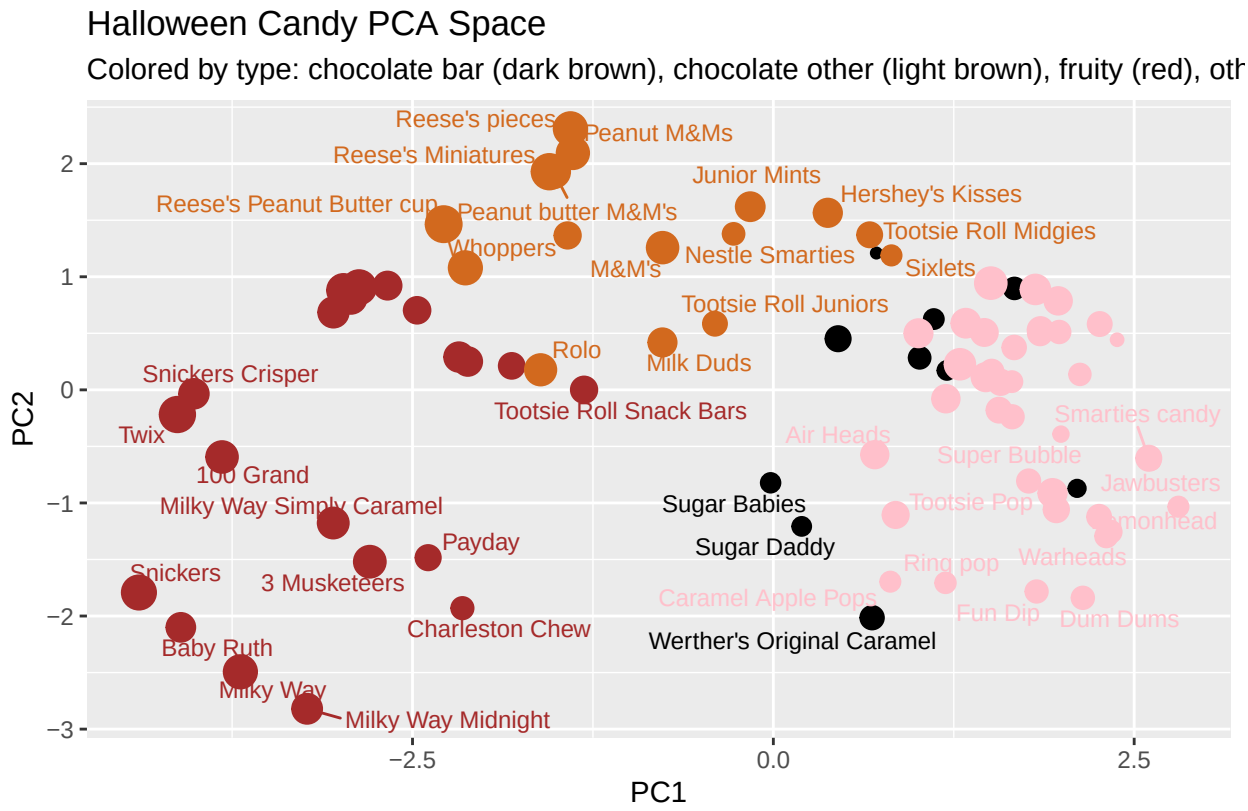

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

```
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
```




```
##relabling plot with nonoverlapping names
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), fruity (re
       caption="Data from 538")
```

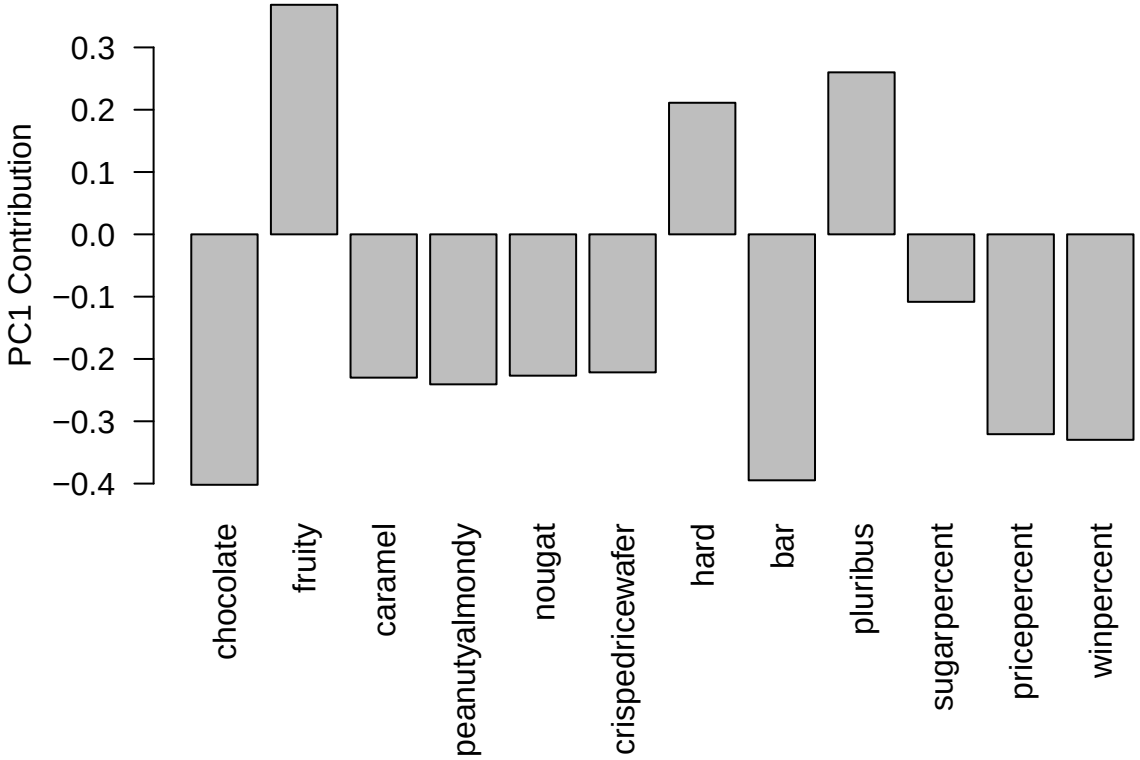
```
## Warning: ggrepel: 44 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Data from 538

```
##making it interactive
##library(plotly)

##ggplotly(p)
```

```
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? Does this make sense?

Variables picked up in the positive direction are fruity, hard, and pluribus. These are common variables for fruity type candy, so yes this does make sense.