class10

Kelsey Fierro

```
candy <- read.csv("candy-data.csv", row.names = 1)
```

```
library(dplyr)
```

Q1. How many different candy types are in this dataset?

```
dim(candy)
```

```
[1] 85 12
```

13 types of candy!

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

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

```
[1] 38
```

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

For M&M's:

```
candy$winpercent[34]
```

```
[1] 66.57458
```

Q4. What is the winpercent value for "Kit Kat"?

```
candy$winpercent[29]
```

```
[1] 76.7686
```

Q5. What is the winpercent value for "Tootsie Roll Snack Bars"?

```
candy$winpercent[78]
```

```
[1] 49.6535
```

Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

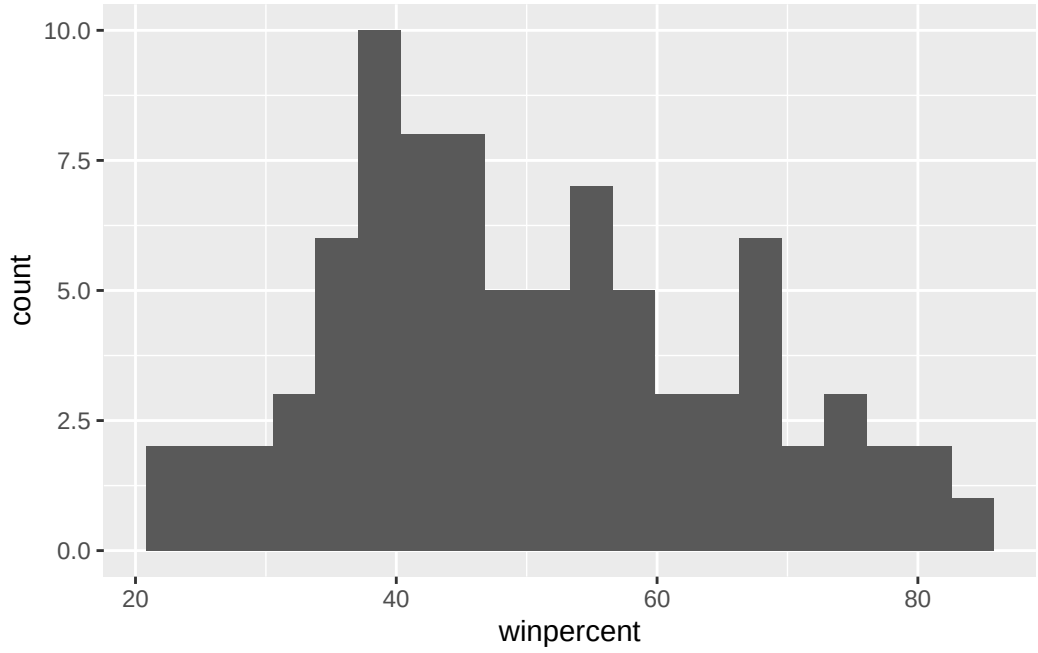
winpercent seems to be the odd column out in terms of scale and measurement.

Q7. What do you think a zero and one represent for the candy$chocolate column?

It represents whether or not the candy has chocolate in it or not

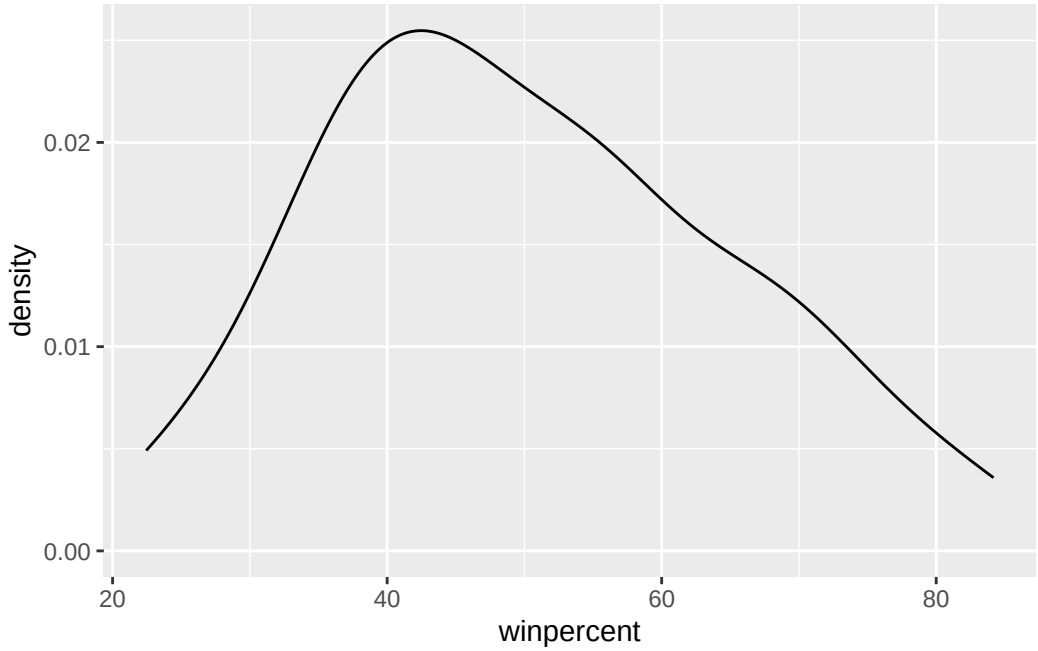
Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical?

```
ggplot(candy) +
  aes(winpercent) +
  geom_density()
```



No the distribution is not symmetrical

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

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

```
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?

```
#find all chocolate candy in dataset, extract their winpercent values, find mean of values, t

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

```
[1] 60.92153
```

```
fruit.inds <- as.logical(candy$fruity)
fruit.candy <- candy[fruit.inds,]
fruit.win <- fruit.candy$winpercent
fruit.mean <- mean(fruit.win)
fruit.mean
```

```
[1] 44.11974
```

```
choc.mean>=fruit.mean
```

```
[1] TRUE
```

Chocolate is ranked higher than fruit

Q12. Is this difference statistically significant?

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

```
    Welch Two Sample t-test

data:  choc.win and fruit.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
```

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

```
ord.ind <- order(candy$winpercent)
head(candy[ord.ind, ], 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?

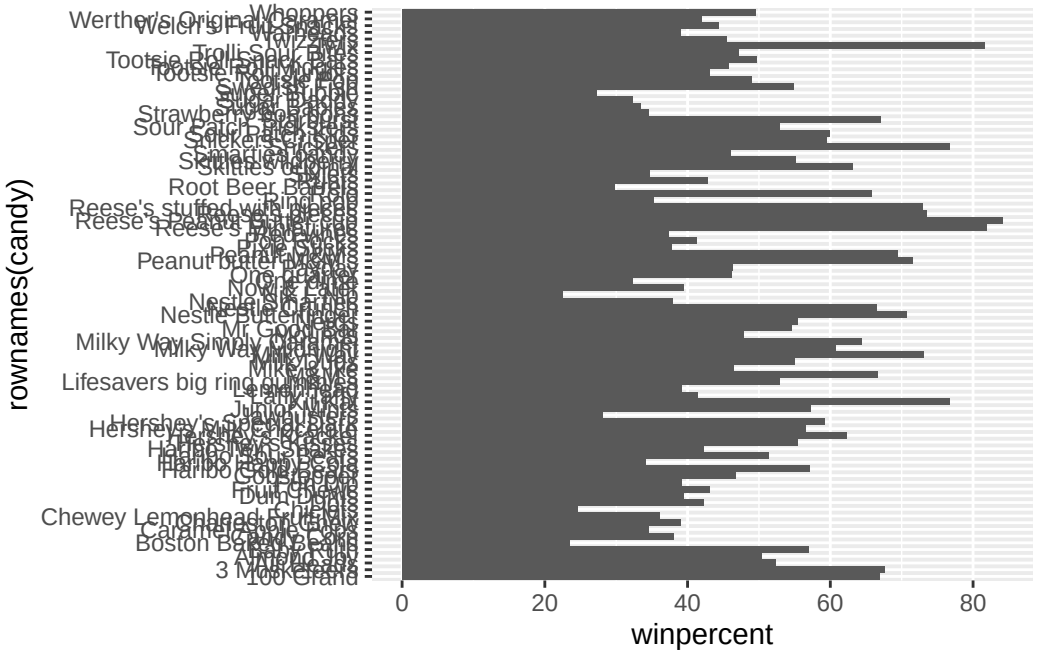
```
tail(candy[ord.ind, ], 5)
```

```
                        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
```

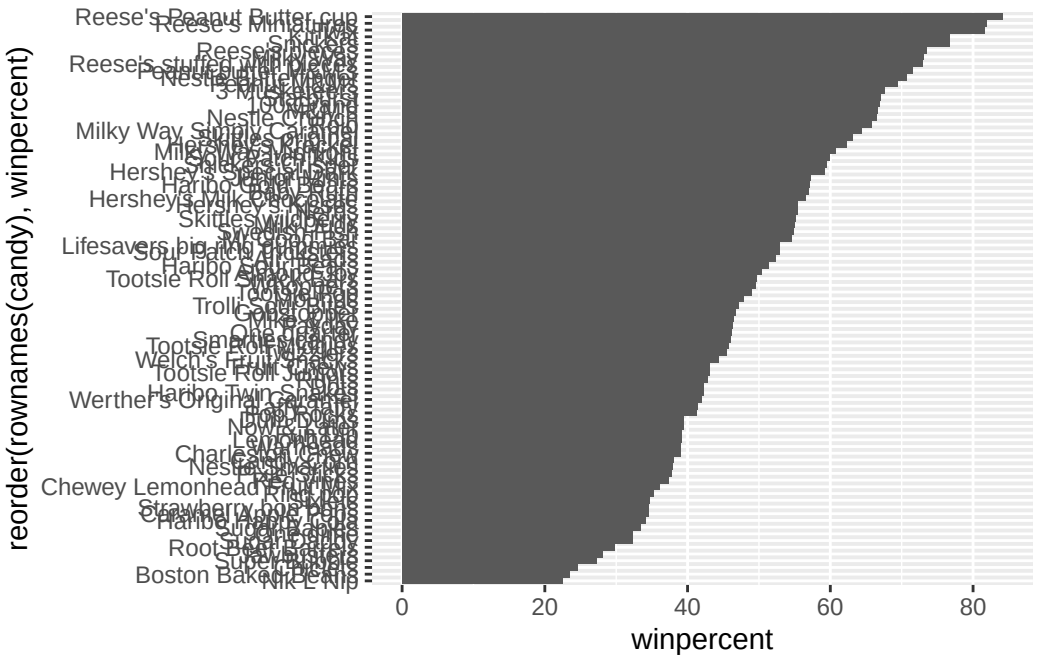
Q15. Make a first barplot of candy ranking based on winpercent values.

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



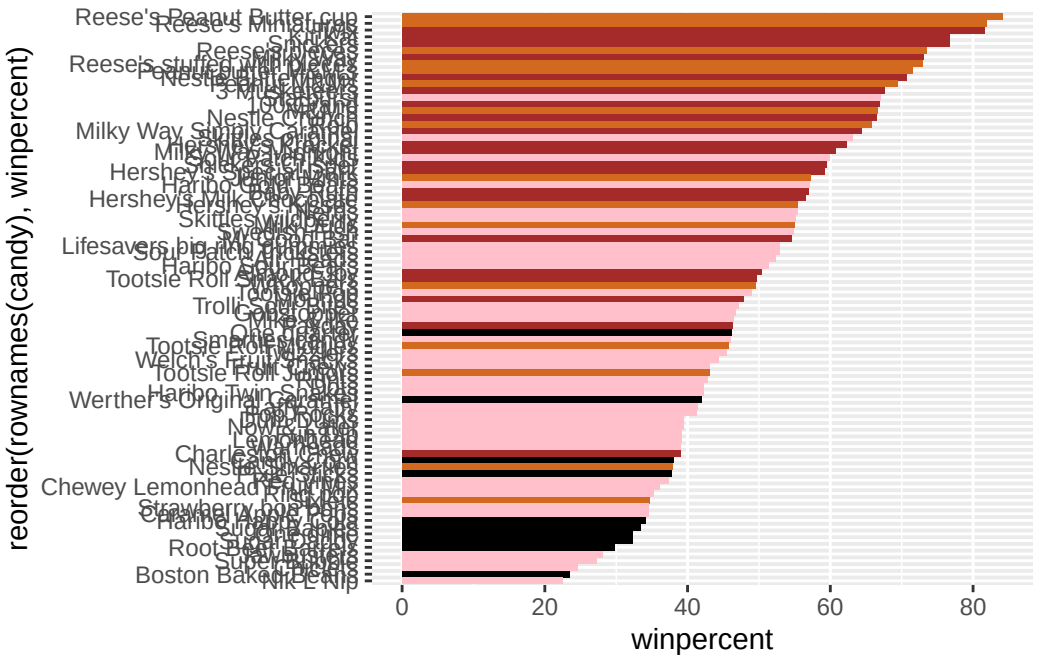
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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




```
mycols <- rep("black", nrow(candy))
mycols[candy$chocolate==1] <- "chocolate"
mycols[candy$bar==1] <- "brown"
mycols[candy$fruity==1] <- "pink"

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



Q17. What is the worst ranked chocolate candy?

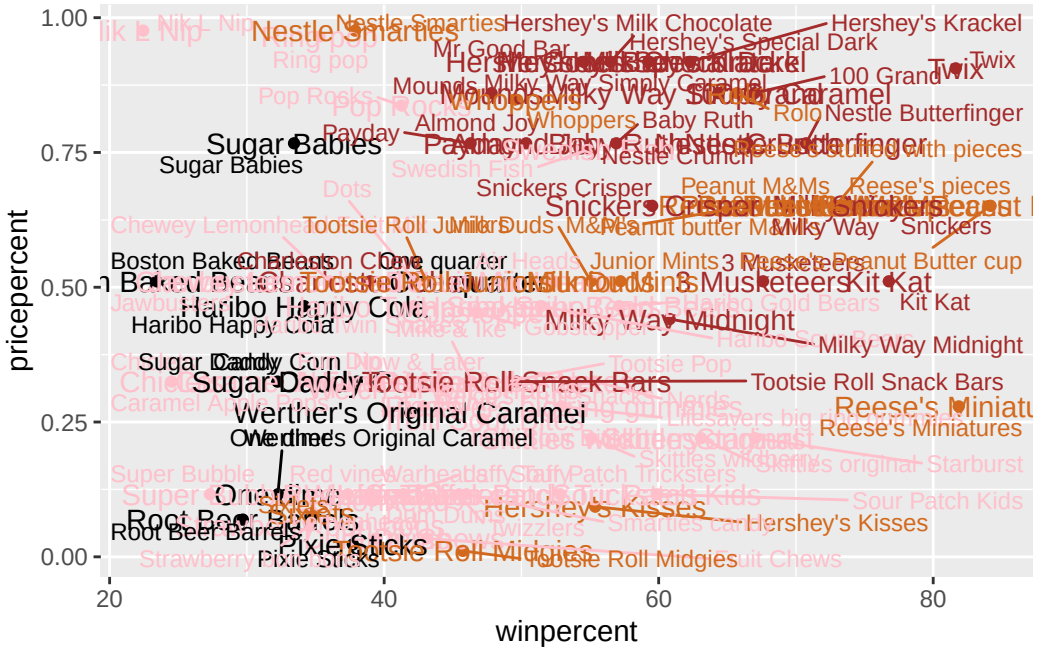
The worst ranked candy is Sixlets.

Q18. What is the best ranked fruity candy?

The best ranked fruity candy is Starburst!

Winpercent vs pricepoint

```
library(ggrepel)
library(ggplot2)
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=mycols) +
  geom_text(col=mycols) +
  geom_text_repel(col=mycols, size=3.3, max.overlaps=50)
```



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?

Reese's miniatures offer the best bang for your buck!

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

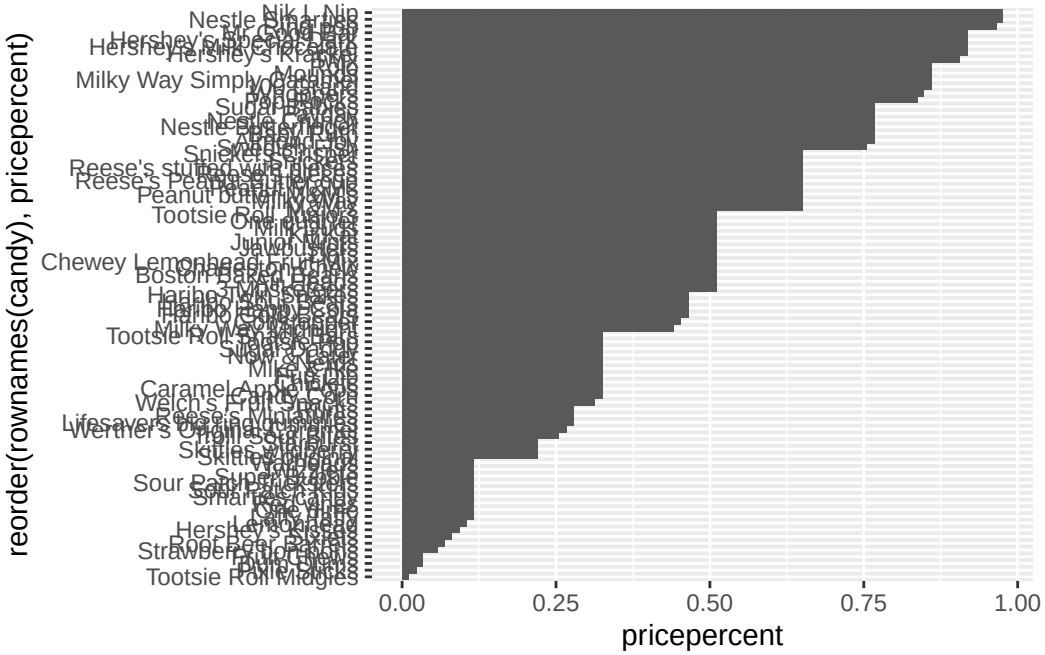
```
ord1 <- order(candy$pricepercent, decreasing = TRUE)
head(candy[ord1,c(11,12)], n=5)
```

```
                      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
```

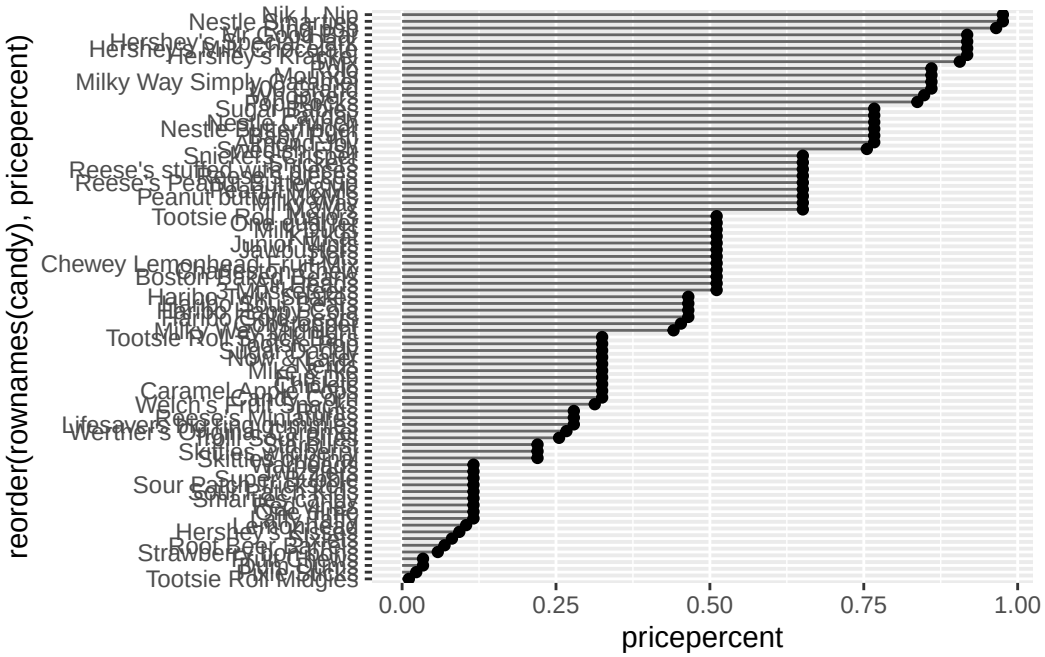
Nik L Nip is tied for most expensive but is somehow one of the least popular candies overall.

Q21. Optional

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




```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
  geom_point()
```



correlation

```
library(corrplot)
```

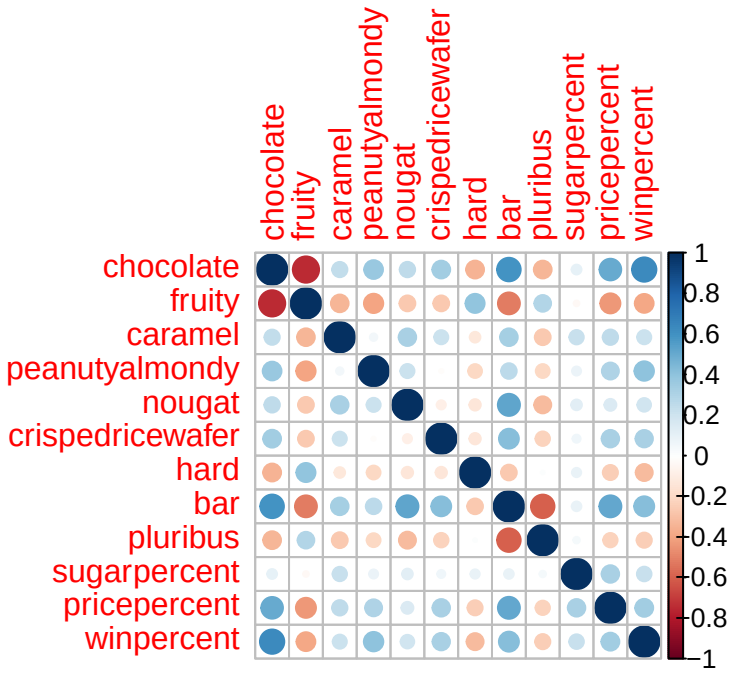
```
corrplot 0.95 loaded
```

```
cor(candy)
```

```
                  chocolate      fruity     caramel peanutyalmondy      nougat
chocolate         1.0000000 -0.74172106  0.24987535     0.37782357  0.25489183
fruity           -0.7417211  1.00000000 -0.33548538    -0.39928014 -0.26936712
caramel           0.2498753 -0.33548538  1.00000000     0.05935614  0.32849280
peanutyalmondy    0.3778236 -0.39928014  0.05935614     1.00000000  0.21311310
nougat            0.2548918 -0.26936712  0.32849280     0.21311310  1.00000000
crispedricewafer  0.3412098 -0.26936712  0.21311310    -0.01764631 -0.08974359
hard             -0.3441769  0.39067750 -0.12235513    -0.20555661 -0.13867505
bar               0.5974211 -0.51506558  0.33396002     0.26041960  0.52297636
pluribus         -0.3396752  0.29972522 -0.26958501    -0.20610932 -0.31033884
sugarpercent      0.1041691 -0.03439296  0.22193335     0.08788927  0.12308135
```

```
pricepercent      0.5046754 -0.43096853  0.25432709     0.30915323  0.15319643
winpercent        0.6365167 -0.38093814  0.21341630     0.40619220  0.19937530
                crispedricewafer        hard         bar    pluribus
chocolate             0.34120978 -0.34417691  0.59742114 -0.33967519
fruity               -0.26936712  0.39067750 -0.51506558  0.29972522
caramel               0.21311310 -0.12235513  0.33396002 -0.26958501
peanutyalmondy       -0.01764631 -0.20555661  0.26041960 -0.20610932
nougat               -0.08974359 -0.13867505  0.52297636 -0.31033884
crispedricewafer      1.00000000 -0.13867505  0.42375093 -0.22469338
hard                 -0.13867505  1.00000000 -0.26516504  0.01453172
bar                   0.42375093 -0.26516504  1.00000000 -0.59340892
pluribus             -0.22469338  0.01453172 -0.59340892  1.00000000
sugarpercent          0.06994969  0.09180975  0.09998516  0.04552282
pricepercent          0.32826539 -0.24436534  0.51840654 -0.22079363
winpercent            0.32467965 -0.31038158  0.42992933 -0.24744787
                sugarpercent pricepercent winpercent
chocolate         0.10416906    0.5046754  0.6365167
fruity           -0.03439296   -0.4309685 -0.3809381
caramel           0.22193335    0.2543271  0.2134163
peanutyalmondy    0.08788927    0.3091532  0.4061922
nougat            0.12308135    0.1531964  0.1993753
crispedricewafer  0.06994969    0.3282654  0.3246797
hard              0.09180975   -0.2443653 -0.3103816
bar               0.09998516    0.5184065  0.4299293
pluribus          0.04552282   -0.2207936 -0.2474479
sugarpercent      1.00000000    0.3297064  0.2291507
pricepercent      0.32970639    1.0000000  0.3453254
winpercent        0.22915066    0.3453254  1.0000000
```

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



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

Chocolate and fruity are anti-correlated, as well as pluribus and bar.

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

Winpercent and chocolate, and bar and chocolate are all positively correlated.

principal component analysis

the main function in base R for this 'prcomp()' and we want to set 'scale=TRUE'

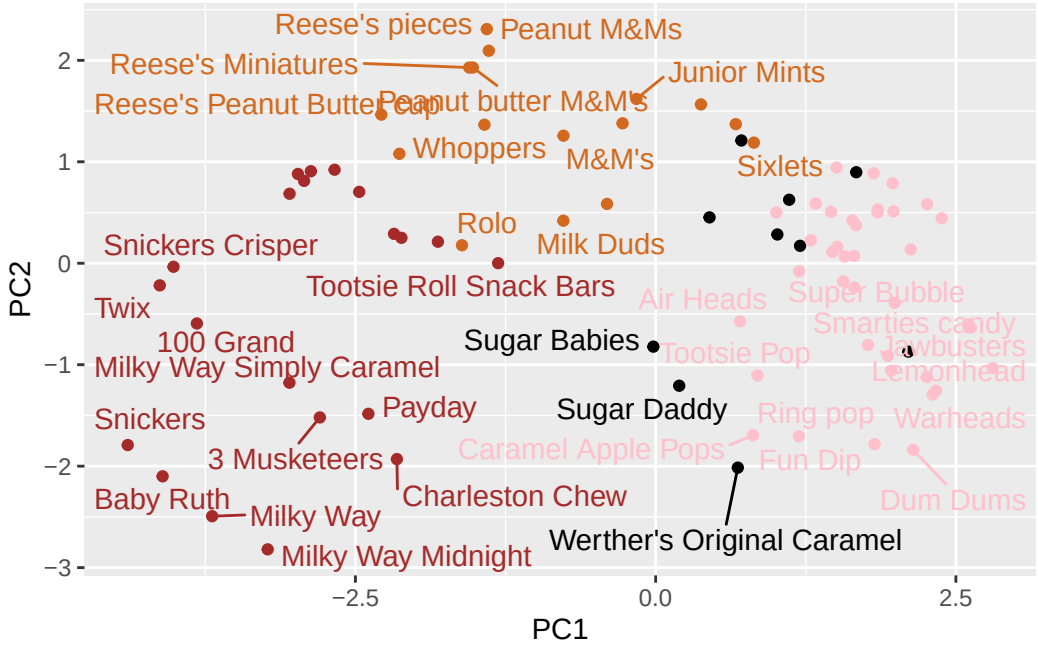
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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
```

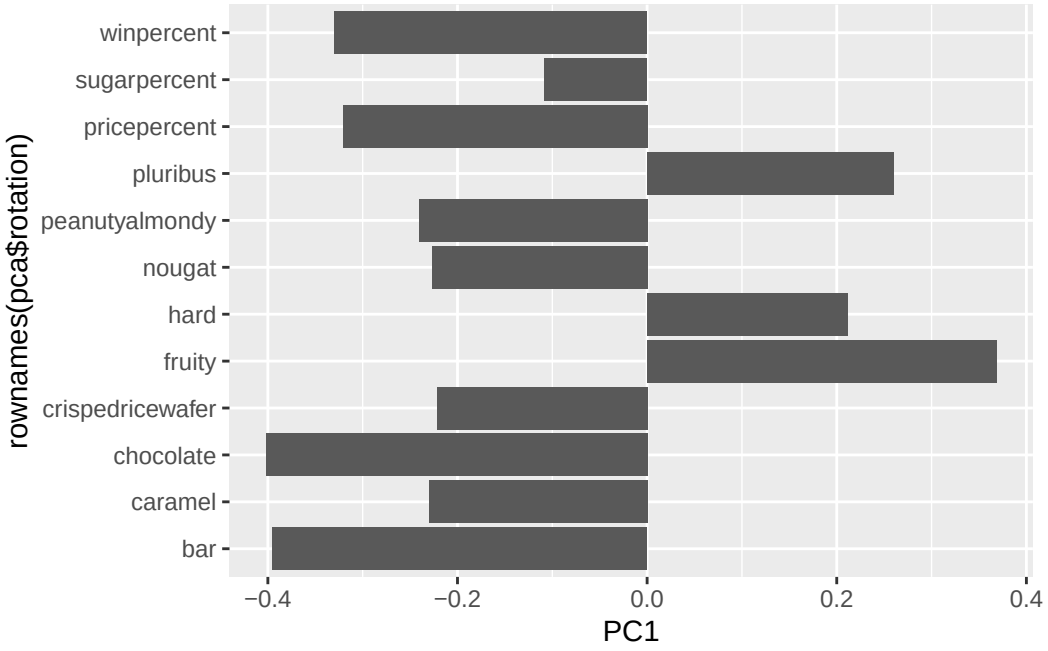
```
#pca$x
library(ggrepel)
ggplot(pca$x) +
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=mycols) +
  geom_text_repel(col=mycols)
```

```
Warning: ggrepel: 48 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



dont forget about your variable "loadings" - how the original variables contribute to your new PCs ...

```
ggplot(pca$rotation) +
  aes(PC1, rownames(pca$rotation)) +
  geom_col()
```



Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

PC1 has the fruity and pluribus variables in a strong positive direction. This makes sense as many fruity candies are small pieces of candy that are packaged together in a bag.