Lab 10: Halloween Candy Mini Project

AUTHOR
Nichelle Camden

Background

In this mini-project we will examine 538 Halloween Candy data.

```
candy <- read.csv("candy-data.txt", 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 this dataset?

```
nrow(candy)
```

```
[1] 85
```

There are 85 different types of candy.

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

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

```
[1] 38
```

There are 38 candy types that are fruity.

Q3. What is your favorite candy in the dataset and what is its winpercent value?

```
#rownames(candy)
candy["Peanut butter M&MÕs",]$winpercent
```

```
[1] 71.46505
```

My favorite candy in the data set is peanut butter M&Ms. Their win percent is 71.4605%

Q4. What is the winpercent value for “Kit Kat”?

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

```
[1] 76.7686
```

The winpercent for Kit Kat is 76.7686%

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

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

```
[1] 49.6535
```

The winpercent for Tootsie Rolls is 49.6535%.

```
#skim(candy)
```

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

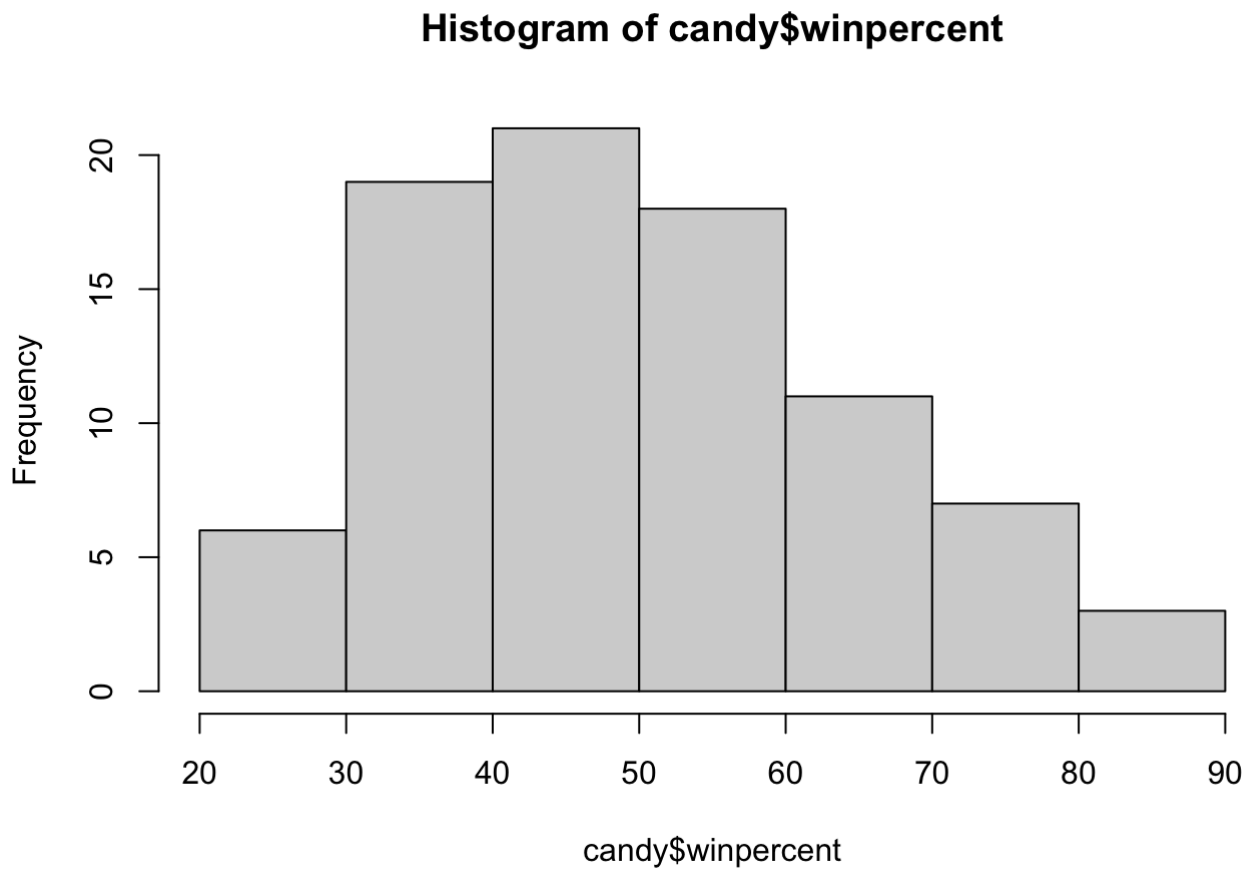
Chocolate

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

0 means no or missing/ that value doesn’t apply here, and 1 means yes or that it does apply.

Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent)
```



Q9. Is the distribution of winpercent values symmetrical?

It is pretty close but the values are a little higher in the first half.

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

below

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

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

higher

Q12. Is this difference statistically significant?

We can then use this logical vector to access the coresponding candy rows (those with TRUE values). For example to get the winpercent values for all nougat contaning candy we can use the code: candy $winpercent[as.logical(candy$nougat)]. In addation the functions mean() and t.test() should help you answer the last two questions here.

```
t.test(candy$winpercent[as.logical(candy$chocolate)],candy$winpercent[as.logical(candy$fr
```

```
    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
```

The p-value is very small so the difference is statistically significant.

3. Overall candy rankings

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

```
inds <- order(candy$winpercent)
head(candy[inds,], 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
```

Q14. What are the top 5 all time favorite candy types out of this set?

```
inds <- order(candy$winpercent)
tail(candy[inds,], n=5)
```

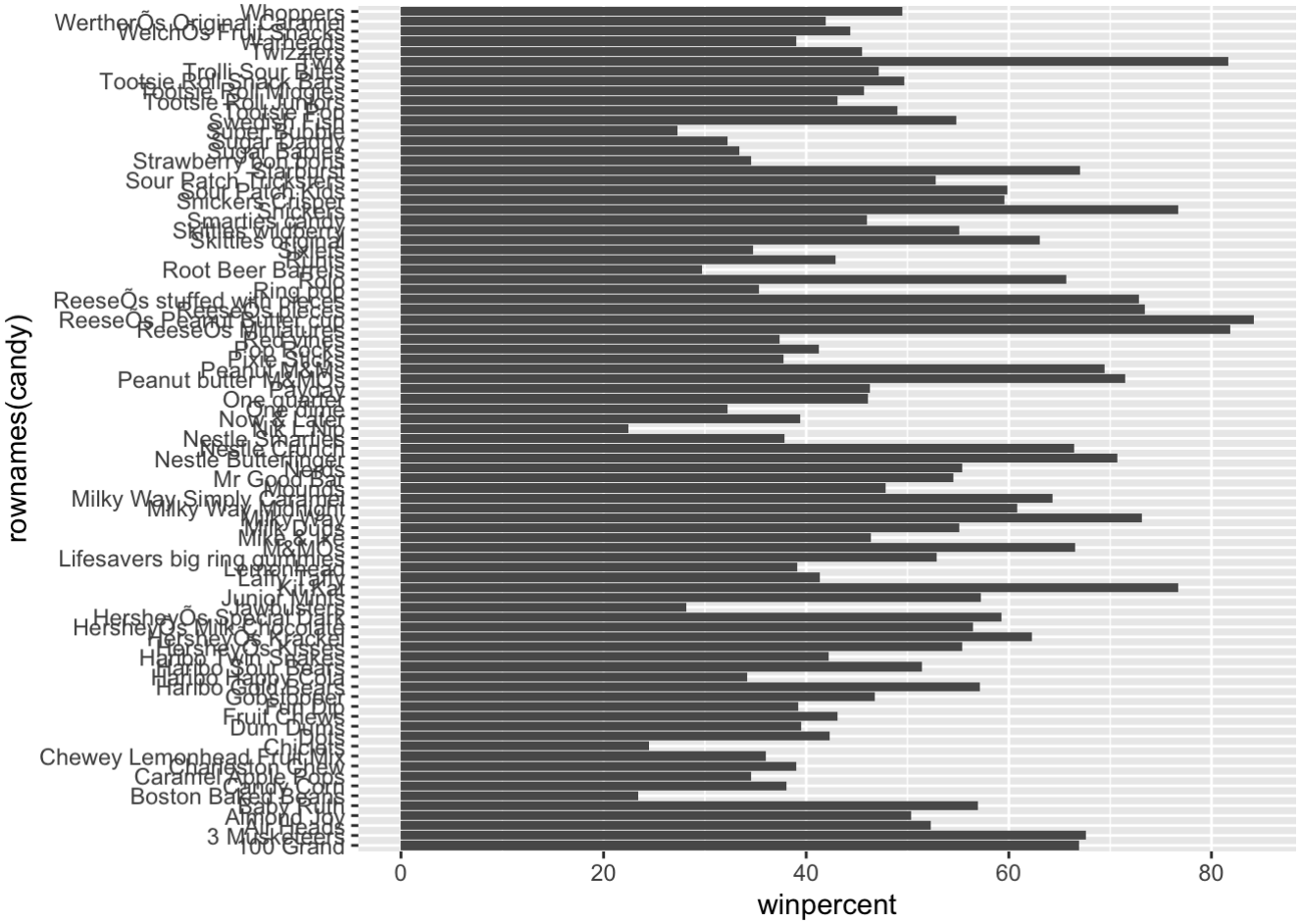
```
                        chocolate fruity caramel peanutyalmondy nougat
Snickers                        1      0       1              1      1
Kit Kat                         1      0       0              0      0
Twix                            1      0       1              0      0
ReeseÕs Miniatures              1      0       0              1      0
ReeseÕs 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
ReeseÕs Miniatures                     0    0   0        0        0.034
ReeseÕs 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
ReeseÕs Miniatures             0.279   81.86626
ReeseÕs Peanut Butter cup      0.651   84.18029
```

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

winpercent values:

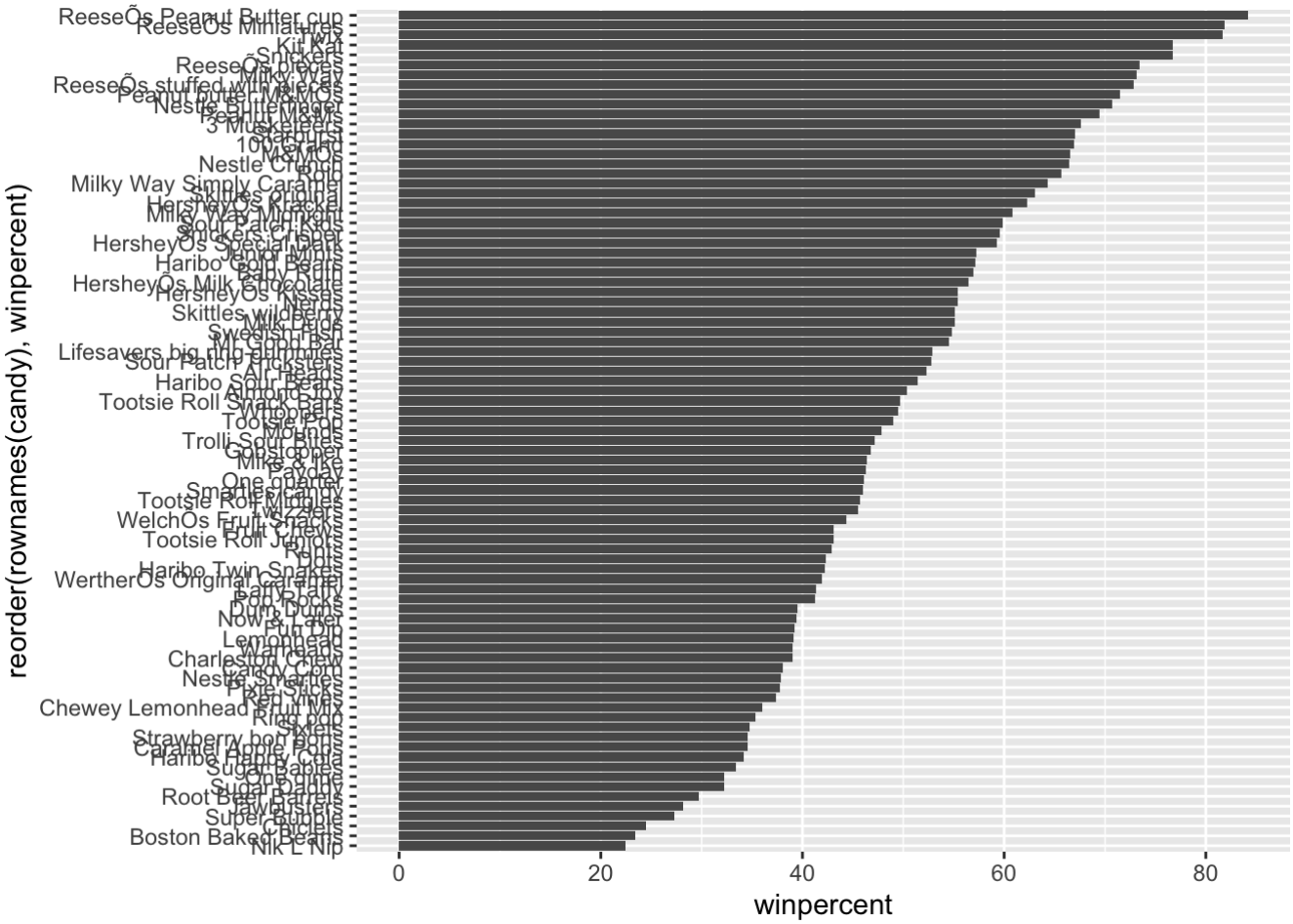
```
library("ggplot2")

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



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

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




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

```
Saving 7 x 5 in image
```

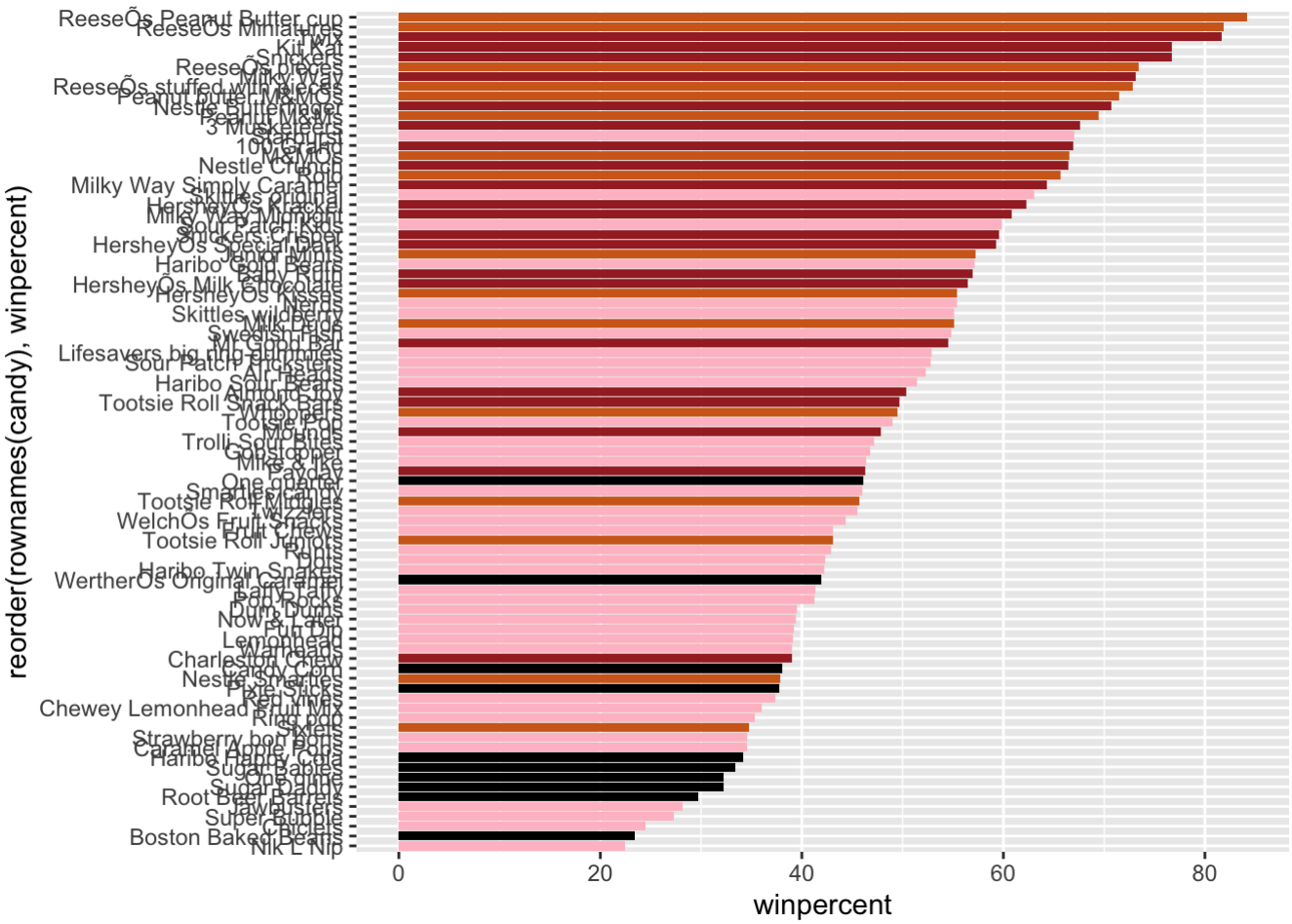
```
my_cols <- rep("black", nrow(candy))
#my_cols
my_cols[as.logical(candy$chocolate)] <- "chocolate"
my_cols[as.logical(candy$bar)] <- "brown"
my_cols[as.logical(candy$fruity)] <- "pink"
my_cols
```

```
 [1] "brown"     "brown"     "black"     "black"     "pink"      "brown"
 [7] "brown"     "black"     "black"     "pink"      "brown"     "pink"
[13] "pink"      "pink"      "pink"      "pink"      "pink"      "pink"
[19] "pink"      "black"     "pink"      "pink"      "chocolate" "brown"
[25] "brown"     "brown"     "pink"      "chocolate" "brown"     "pink"
[31] "pink"      "pink"      "chocolate" "chocolate" "pink"      "chocolate"
[37] "brown"     "brown"     "brown"     "brown"     "brown"     "pink"
[43] "brown"     "brown"     "pink"      "pink"      "brown"     "chocolate"
[49] "black"     "pink"      "pink"      "chocolate" "chocolate" "chocolate"
[55] "chocolate" "pink"      "chocolate" "black"     "pink"      "chocolate"
[61] "pink"      "pink"      "chocolate" "pink"      "brown"     "brown"
[67] "pink"      "pink"      "pink"      "pink"      "black"     "black"
```

```
[67] "pink"      "pink"      "pink"      "pink"      "black"     "black"
[73] "pink"      "pink"      "pink"      "chocolate" "chocolate" "brown"
[79] "pink"      "brown"     "pink"      "pink"      "pink"      "black"
[85] "chocolate"
```

Now I can use this vecotor to color my barplot

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



Q17. What is the worst ranked chocolate candy?

Sixlets is worst ranked

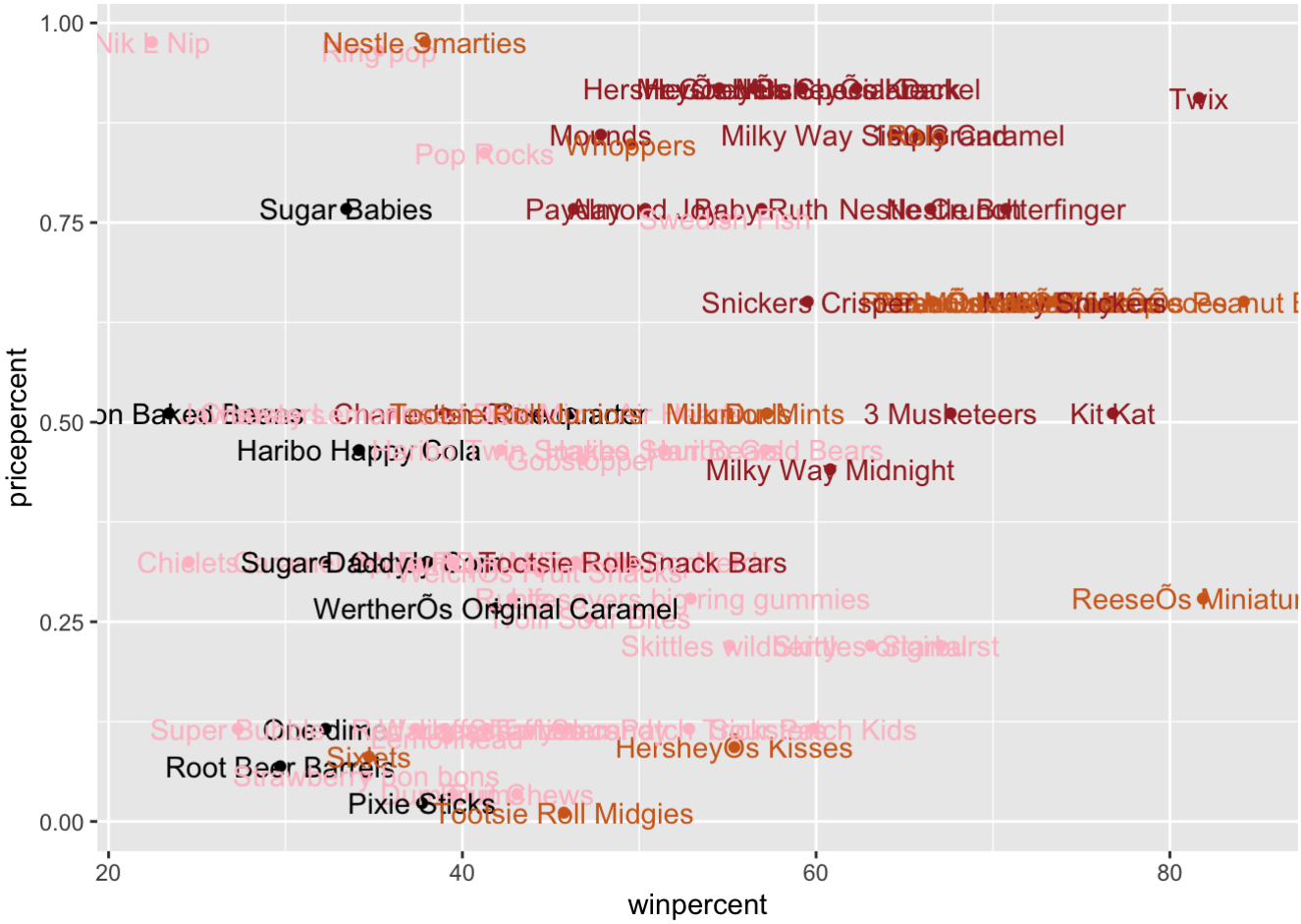
Q18. What is the best ranked fruity candy?

Starburst

4. Taking a look at pricepoint.

What is the best candy for the least amount fo money? One way to get this would be to make a plot of `winpercent` vs `pricepercent` variable.

```
ggplot(candy) +
  aes(winpercent, pricepercent, label = rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text(col=my_cols)
```

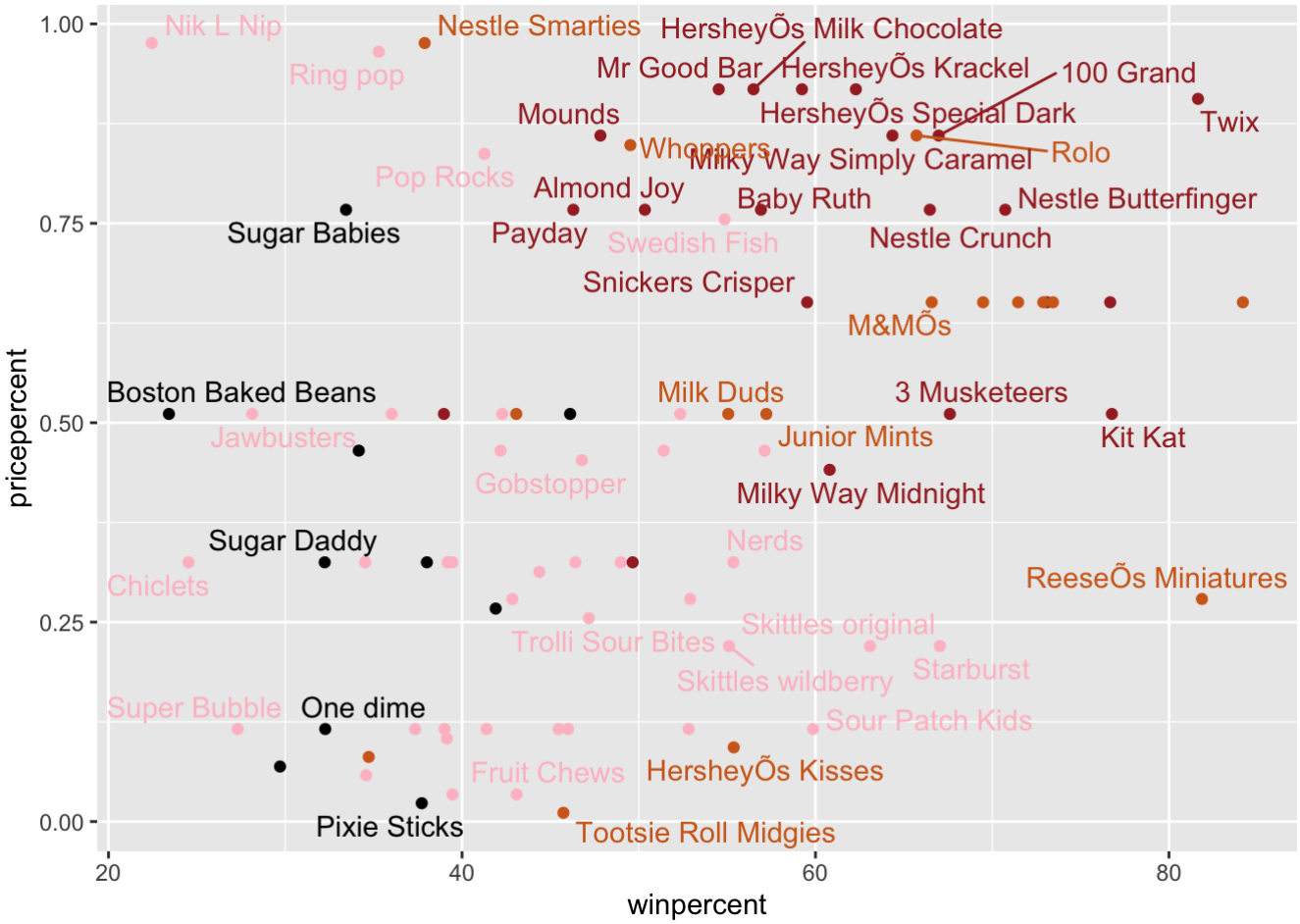


This plot sucks. It's too hard to read the labels. We can use `ggrepel` package to help with this.

```
library("ggrepel")

ggplot(candy) +
  aes(winpercent, pricepercent, label = rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, max.overlaps =8)
```

```
Warning: ggrepel: 39 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



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?

Reeses Miniatures

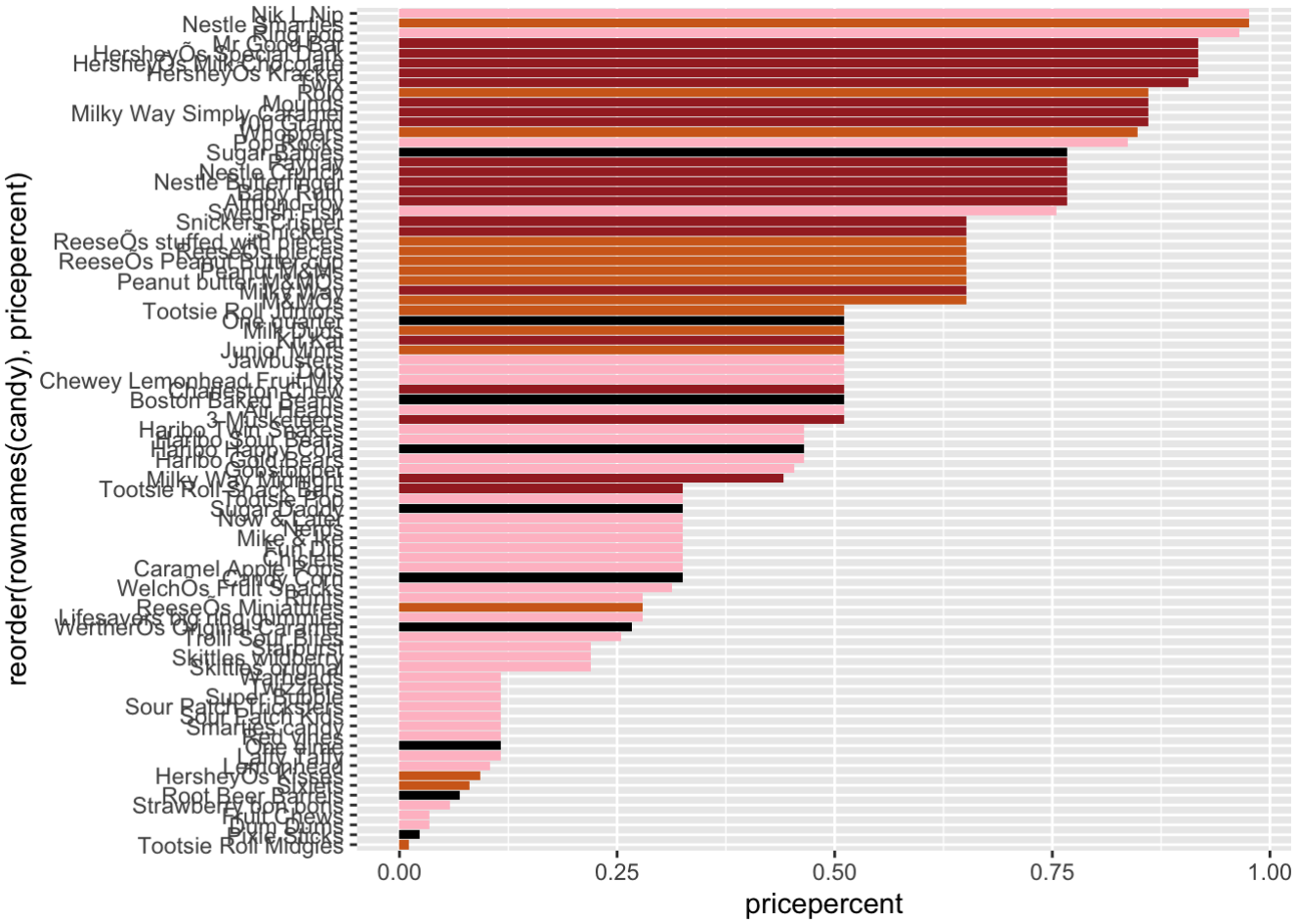
Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

Nik L Nip, Smarties, ring pop, Mr.Good Bar, Hershey's special dark

```
order(candy$pricepercent)
```

```
 [1] 77 49 15 16 70 58 60 23 31  3 30 51 64 67 68 73 81 82 61 62 69 79 84 32 52
[26] 59 83  9 10 13 17 35 42 46 72 75 78 38 18 19 20 21 22  2  4  5  8 11 12 14
[51] 27 28 29 36 76 33 34 37 48 53 54 55 65 66 74  6  7 43 44 47 71 50 85  1 39
[76] 40 57 80 24 25 26 41 56 45 63
```

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

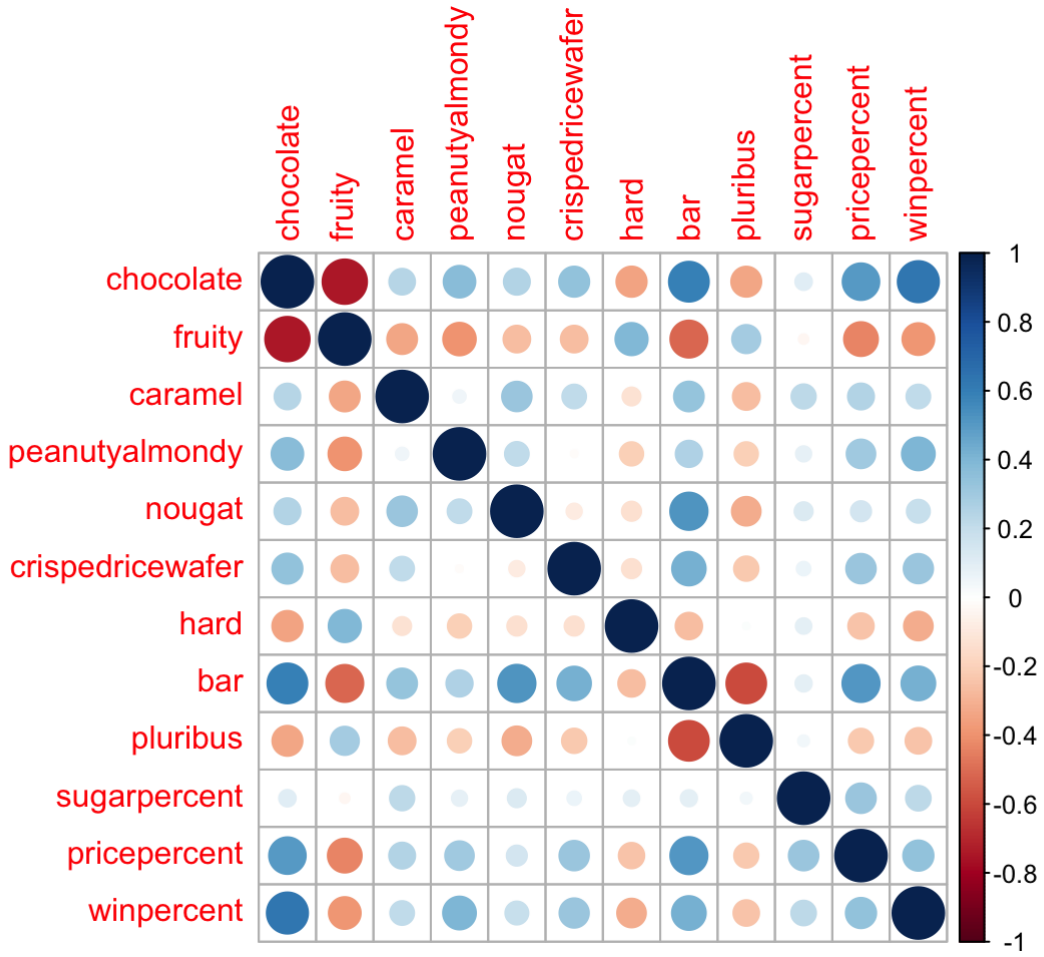


5. Exploring the correlation structure

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



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

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

chocolate and bar or chocolate and winpercent

PCA: Principal Component Analysis

The main function that's always there for us is `prcomp`. It has an important argument that is set to `scale=FALSE`.

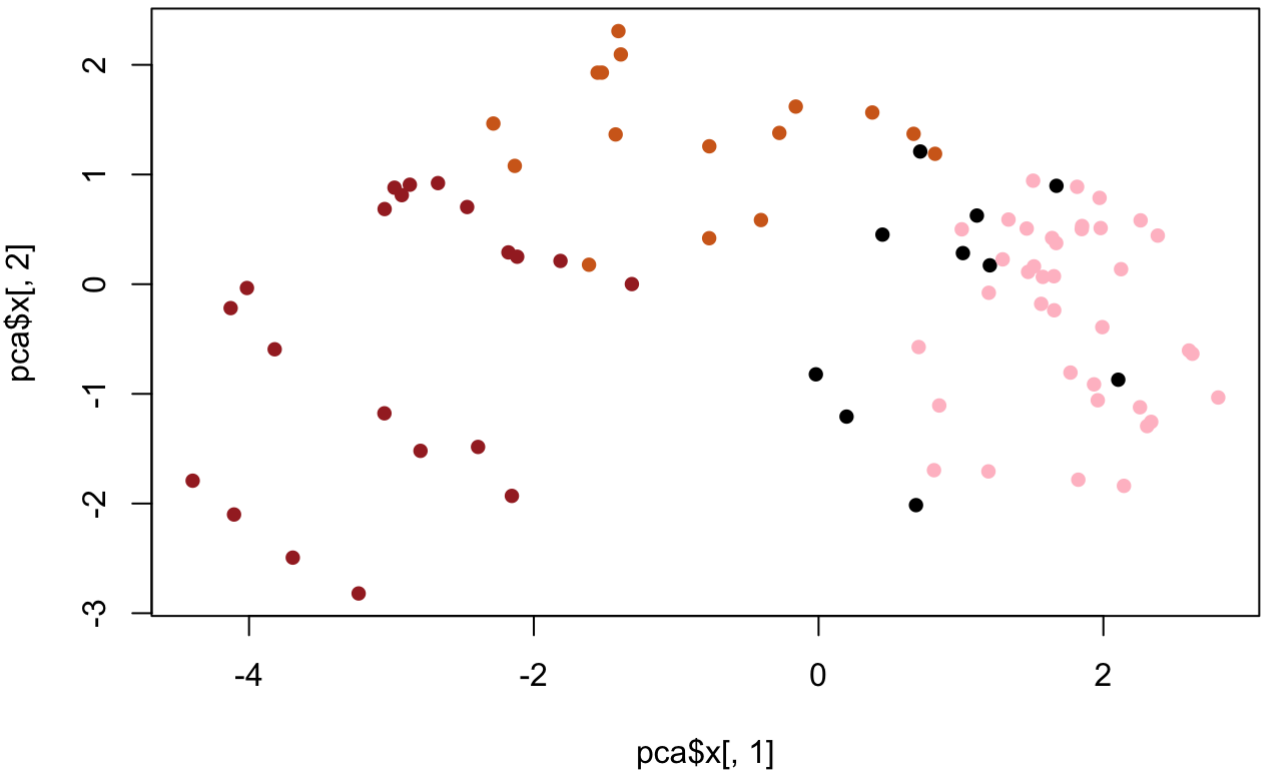
```
pca <- prcomp(candy, scale =TRUE)
summary(pca)
```

```
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
```

My PCA plot (aka PC1 vs PC2) score plot

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

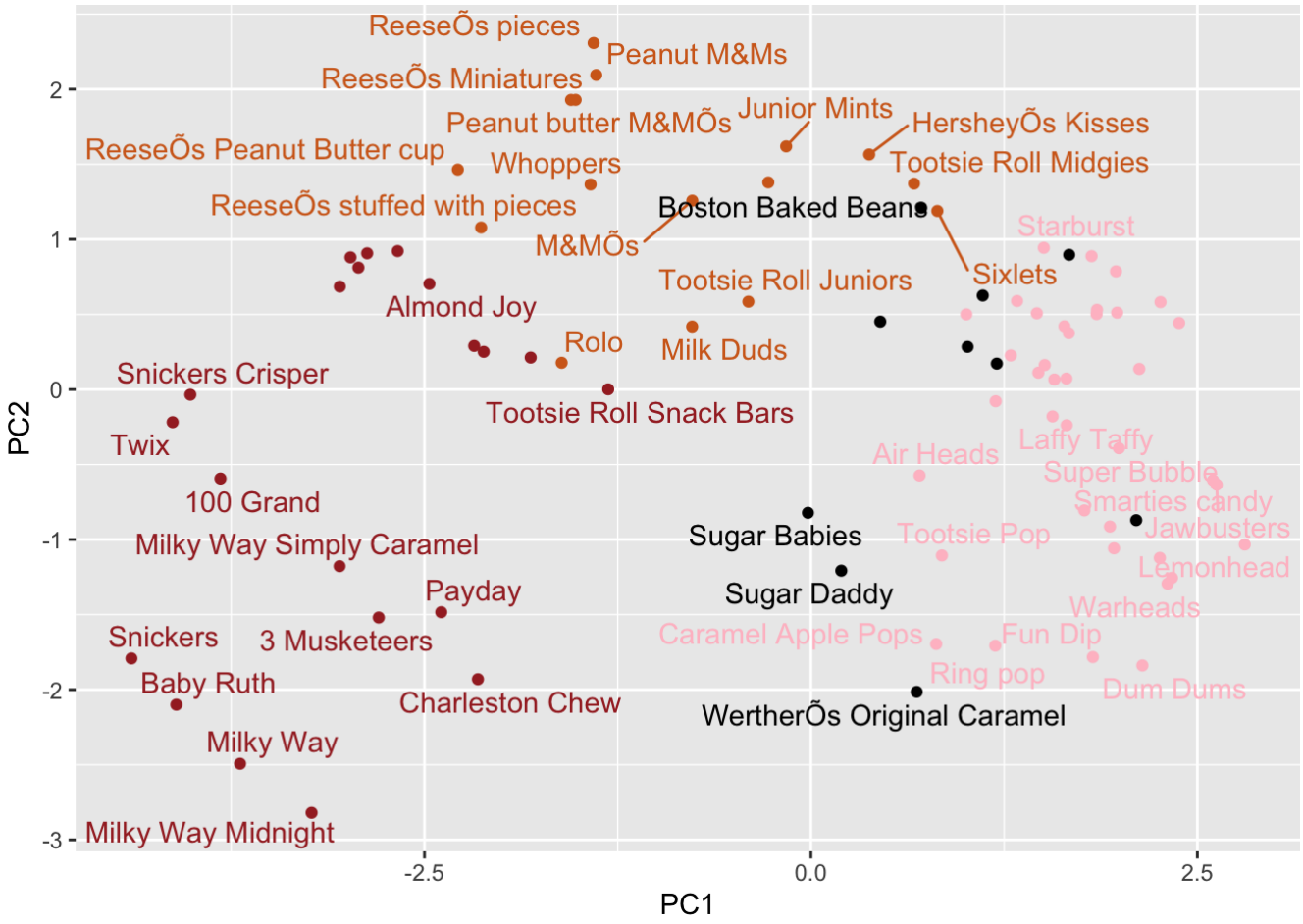


I will make a nicer plot with ggplot. ggplot only works with data.frames as input so I need to make one for it first..

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

```
ggplot(my_data) +
  aes(PC1, PC2, labels = rownames(my_data)) +
  geom_point(col= my_cols) +
  geom_text_repel(label= rownames(my_data), col=my_cols, max.overlaps = 7)
```

```
Warning: ggrepel: 40 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```


ReeseÕs pieces
Peanut M&Ms
ReeseÕs Miniatures
Peanut butter M&MÕs
Junior Mints
HersheyÕs Kisses
ReeseÕs Peanut Butter cup
Whoppers
Tootsie Roll Midgies
ReeseÕs stuffed with pieces
Boston Baked Beans
Starburst
M&MÕs
Sixlets
Tootsie Roll Juniors
Almond Joy
Rolo
Milk Duds
Snickers Crisper
Tootsie Roll Snack Bars
Laffy Taffy
Air Heads
Twix
Super Bubble
100 Grand
Smarties candy
Jawbusters
Sugar Babies
Tootsie Pop
Milky Way Simply Caramel
Lemonhead
Payday
Sugar Daddy
Warheads
Snickers
3 Musketeers
Caramel Apple Pops
Fun Dip
Baby Ruth
Ring pop
Charleston Chew
Dum Dums
WertherÕs Original Caramel
Milky Way
Milky Way Midnight
PC2
PC1
2
1
0
-1
-2
-3
-2.5
0.0
2.5