lab10

1. Importing candy data

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

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 are in this dataset?

85

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

38

> **Q** What are these fruity candy

```r
rownames(candy[candy$fruity == 1, ])
```

```
 [1] "Air Heads"                   "Caramel Apple Pops"
 [3] "Chewey Lemonhead Fruit Mix"  "Chiclets"
 [5] "Dots"                        "Dum Dums"
 [7] "Fruit Chews"                 "Fun Dip"
 [9] "Gobstopper"                  "Haribo Gold Bears"
[11] "Haribo Sour Bears"           "Haribo Twin Snakes"
[13] "Jawbusters"                  "Laffy Taffy"
[15] "Lemonhead"                   "Lifesavers big ring gummies"
[17] "Mike & Ike"                  "Nerds"
[19] "Nik L Nip"                   "Now & Later"
[21] "Pop Rocks"                   "Red vines"
[23] "Ring pop"                    "Runts"
[25] "Skittles original"           "Skittles wildberry"
[27] "Smarties candy"              "Sour Patch Kids"
[29] "Sour Patch Tricksters"       "Starburst"
[31] "Strawberry bon bons"         "Super Bubble"
[33] "Swedish Fish"                "Tootsie Pop"
[35] "Trolli Sour Bites"           "Twizzlers"
[37] "Warheads"                    "Welch's Fruit Snacks"
```

2. What is your favorite candy?

How often does my favorite candy win?

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

```
[1] 81.64291
```

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

```
[1] 81.64291
```

```r
# install.packages("skimr")
# skimr::skim(candy) # if you only want to use this one funciton from this pacakge
library("skimr")
```

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

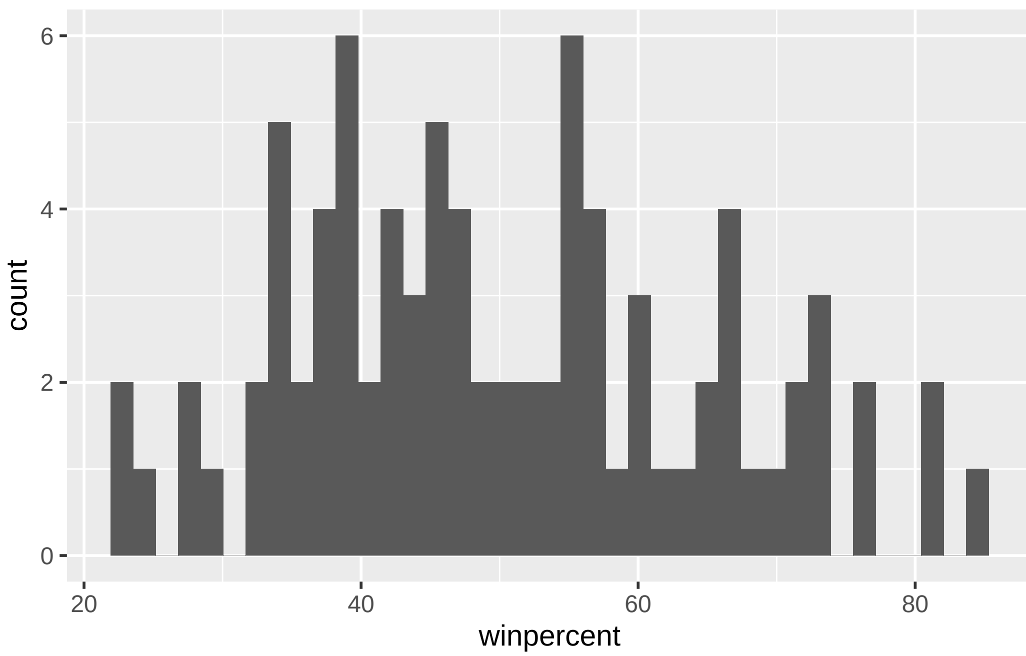
Yes. Winpercent column has mean that's a few times bigger than the means of other columns.

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

Zero means the candy is not classified as containing chocolate.

> **Q8**. Plot a histogram of winpercent values

```r
library(ggplot2)
ggplot(candy) +
  aes(x = winpercent) +
  geom_histogram(bins = 39)
```



Q9. Is the distribution of winpercent values symmetrical?

No

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

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

```
[1] 50.31676
```

Above

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

I need to “subset” (a.k.a. “select”, “filter”) to just chocolate candy, get their winpercent values, and then calculate the means of these.

```r
# Filter/select/subset to just chocolate rows
chocolate.candy <- candy[as.logical(candy$chocolate), ]

# Get their winpercent values
chocolate.winpercent <- chocolate.candy$winpercent

# Calculate their mean winpercent value
mean(chocolate.winpercent)
```

```
[1] 60.92153
```

```r
mean(candy[candy$chocolate == 1, ]$winpercent)
```

```
[1] 60.92153
```

```r
mean(candy[candy$fruity == 1, ]$winpercent)
```

```
[1] 44.11974
```

Higher

> **Q12**. Is this difference statistically significant?

```r
chocolate <- candy[candy$chocolate == 1, ]$winpercent
fruity <- candy[candy$fruity == 1, ]$winpercent
t_test <- t.test(chocolate, fruity, alternative = c("greater")) # is chocolate's mean stat
```

P-value $1.4356889 \times 10^{-8} < 0.05$, so statistically significant

3. Overall Candy Rankings

`order()` returns the "indices" of the input that would result in it being sorted.

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

```r
# sort(candy$winpercent, decreasing=FALSE)
rownames(candy[order(candy$winpercent, decreasing=FALSE), ][1:5, ])
```

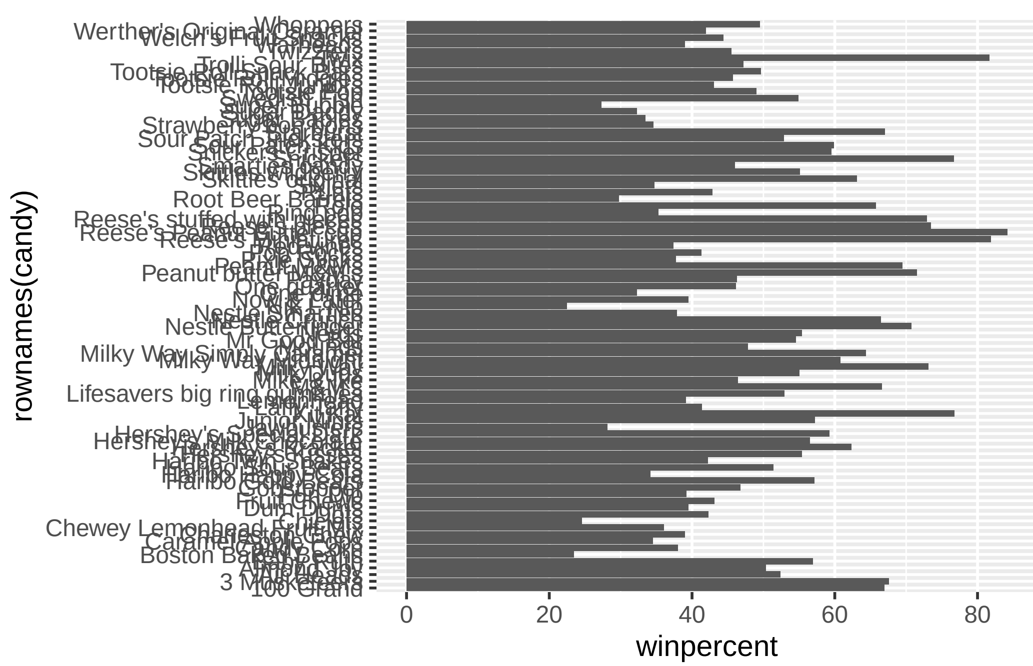
```
[1] "Nik L Nip"          "Boston Baked Beans" "Chiclets"
[4] "Super Bubble"       "Jawbusters"
```

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

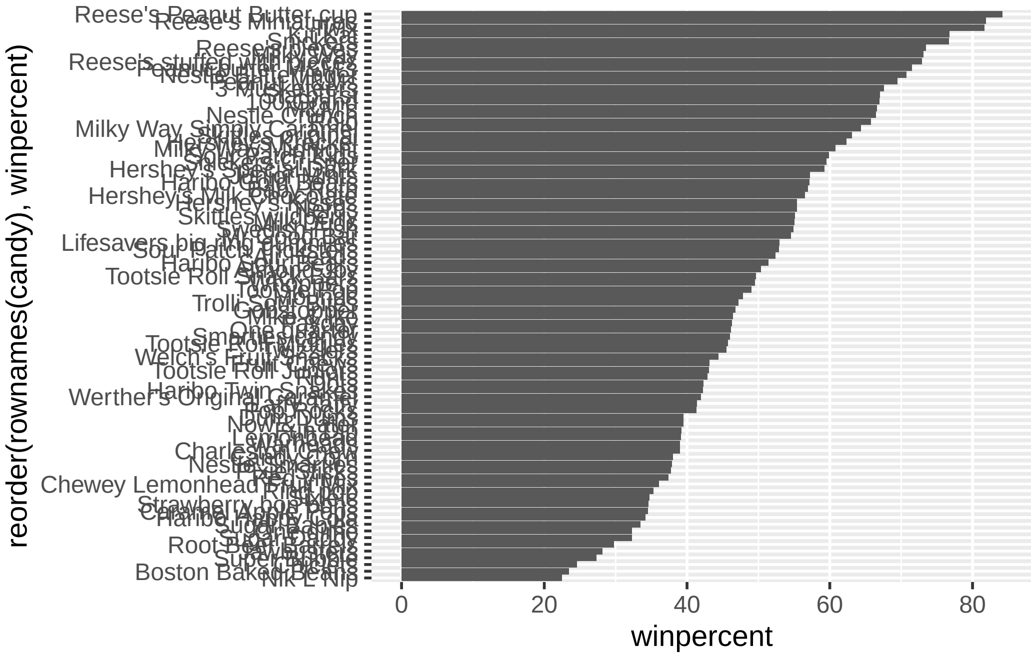
```
rownames(candy[order(candy$winpercent, decreasing=TRUE), ][1:5, ])
```

```
[1] "Reese's Peanut Butter cup" "Reese's Miniatures"
[3] "Twix"                      "Kit Kat"
[5] "Snickers"
```

```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_bar(stat = "identity")
```

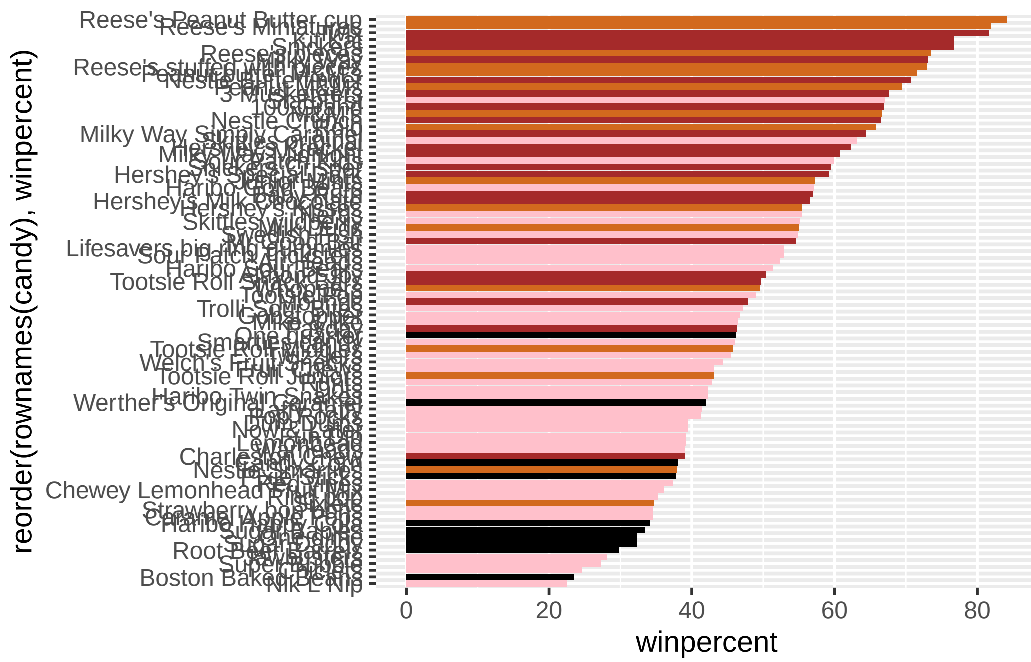



```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_bar(stat = "identity")
```



Time to add some useful color

```
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. What is the worst ranked chocolate candy?

Sixlets

Q18. What is the best ranked fruity candy?

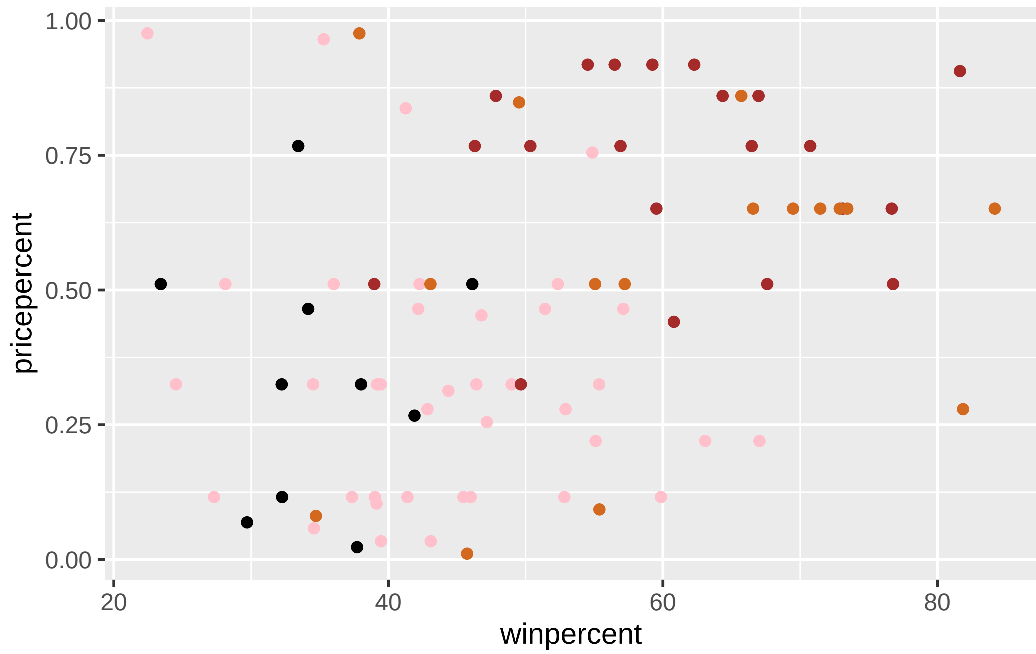
Starburst

4. Taking a look at pricepercent

```r
library(ggrepel)

# How about a plot of price vs win
my_cols[as.logical(candy$fruity)] == "red"
```

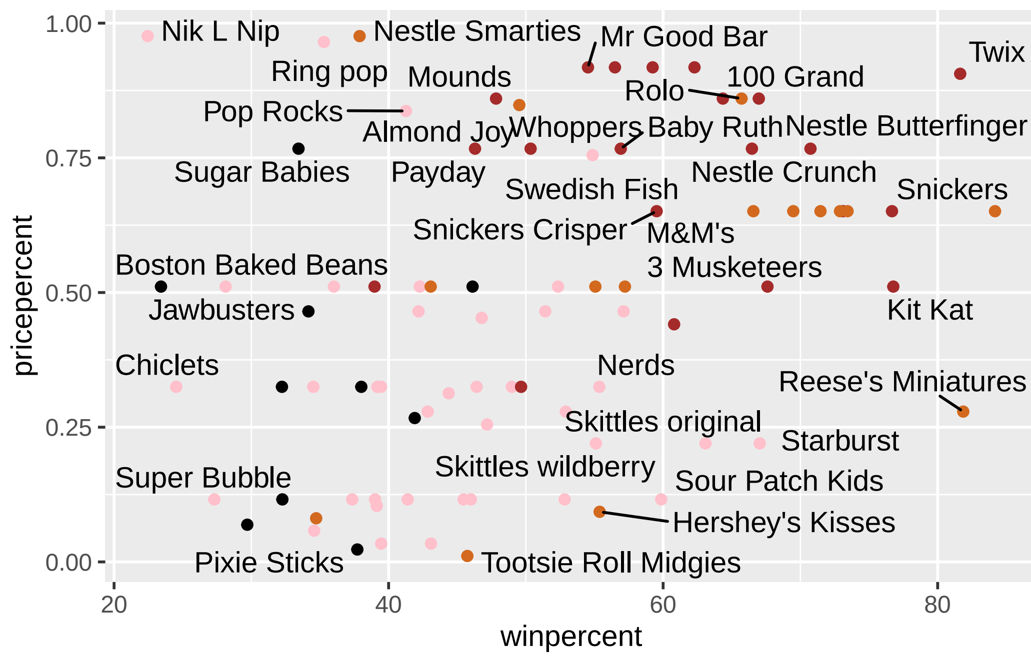
```
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[37] FALSE FALSE
```

```r
ggplot(candy) +
  aes(winpercent, pricepercent) +
  geom_point(col=my_cols)
```



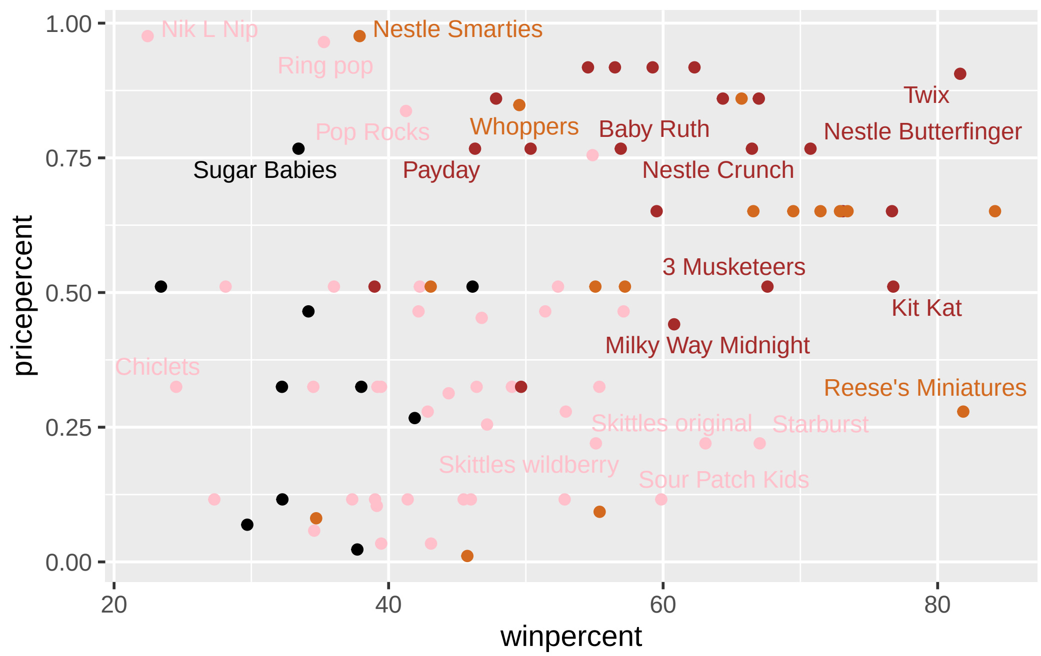

```r
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel()
```

```
Warning: ggrepel: 50 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




```
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**. 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 minatures

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

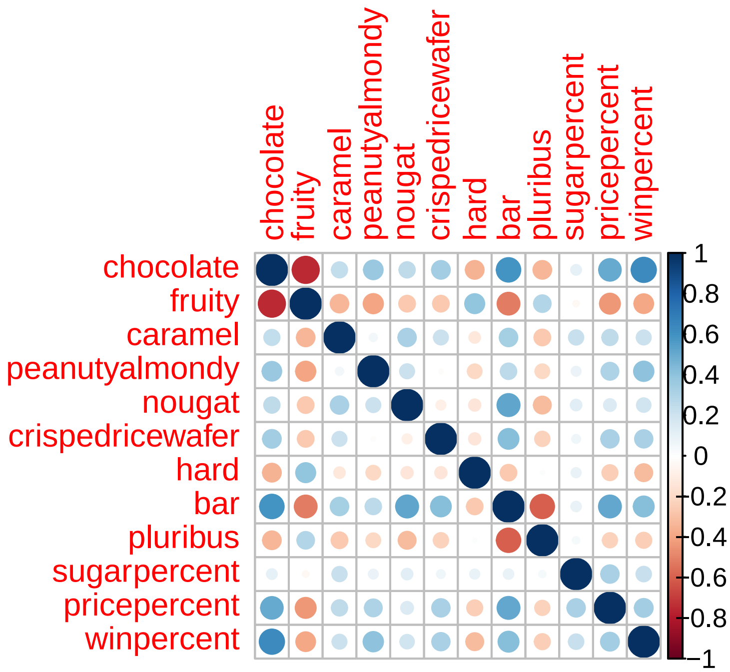
Nik L Nip

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?

winpercent and chocolate

6. Principal Component Analysis

Let's apply PCA using the `prcom()` function to our candy dataset remembering to set the `scale=TRUE/FALSE` argument.

```
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
```

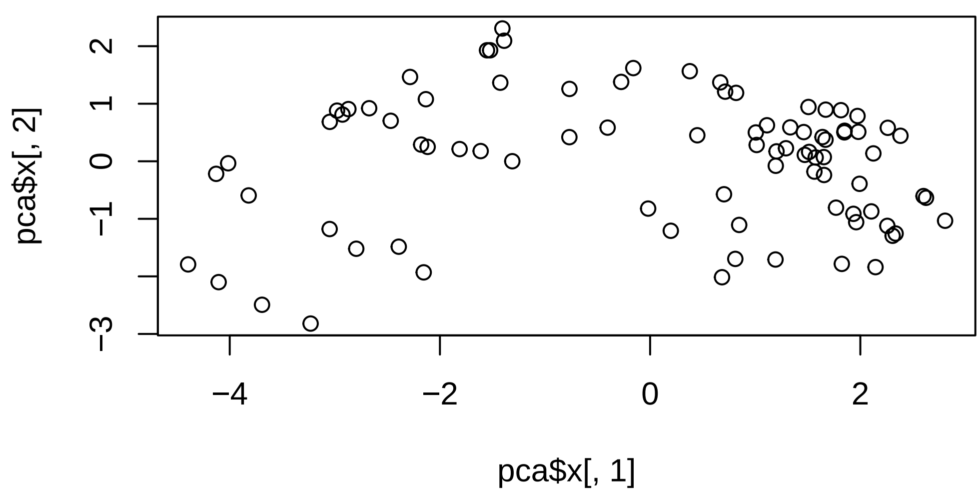
```
                           PC8     PC9    PC10    PC11    PC12
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

```r
pca_unscaled <- prcomp(candy, scale=FALSE)
summary(pca_unscaled)
```

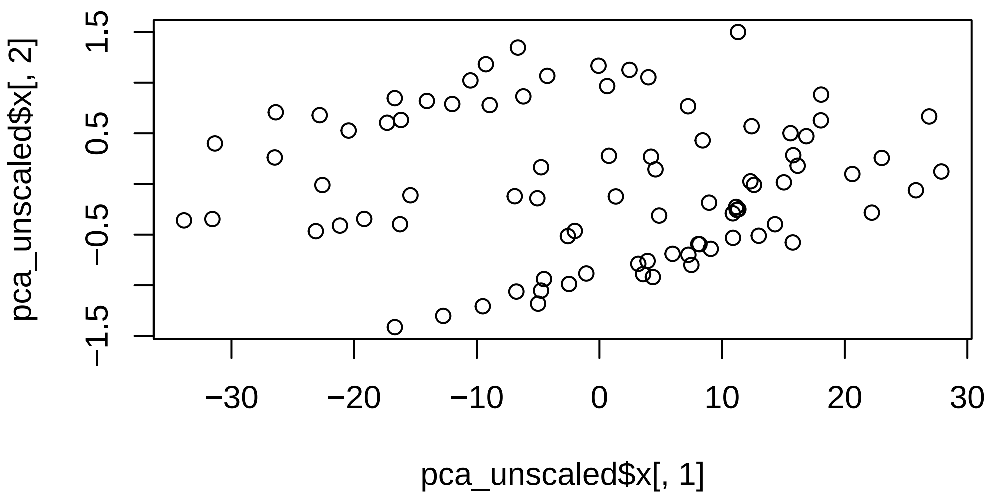
```
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     14.7231 0.70241 0.47762 0.37292 0.34641 0.33614 0.30748
Proportion of Variance  0.9935 0.00226 0.00105 0.00064 0.00055 0.00052 0.00043
Cumulative Proportion   0.9935 0.99574 0.99678 0.99742 0.99797 0.99849 0.99892
                           PC8     PC9    PC10    PC11    PC12
Standard deviation     0.27417 0.23826 0.21435 0.18434 0.15331
Proportion of Variance 0.00034 0.00026 0.00021 0.00016 0.00011
Cumulative Proportion  0.99927 0.99953 0.99974 0.99989 1.00000
```

Now we can plot our main PCA score plot of PC1 vs PC2.

```r
plot(pca$x[, 1], pca$x[, 2])
```

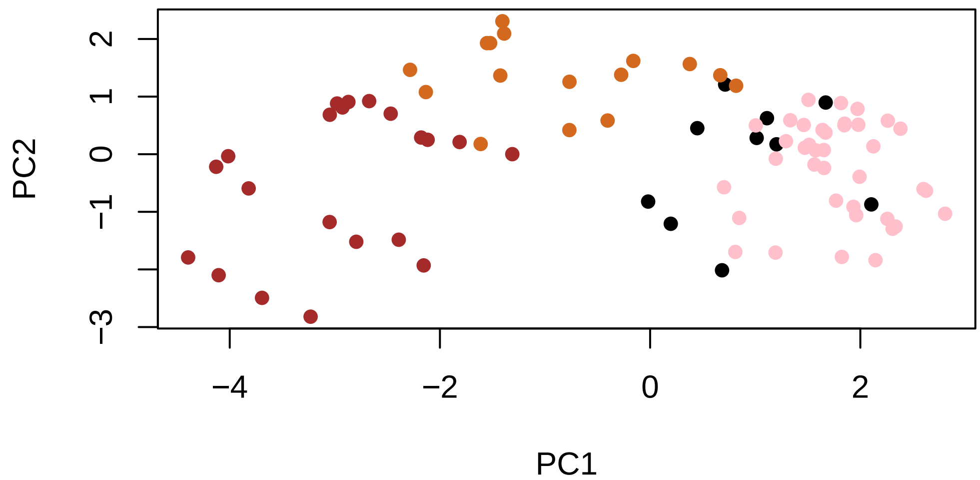


```r
plot(pca_unscaled$x[, 1], pca_unscaled$x[, 2])
```



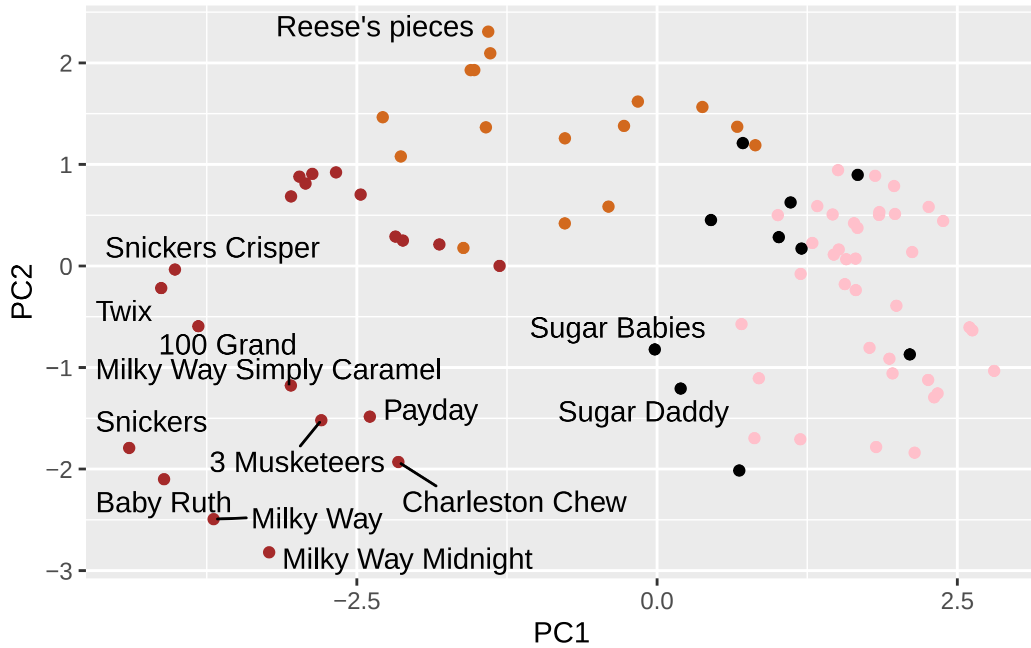
We can change the plotting character and add some color:

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



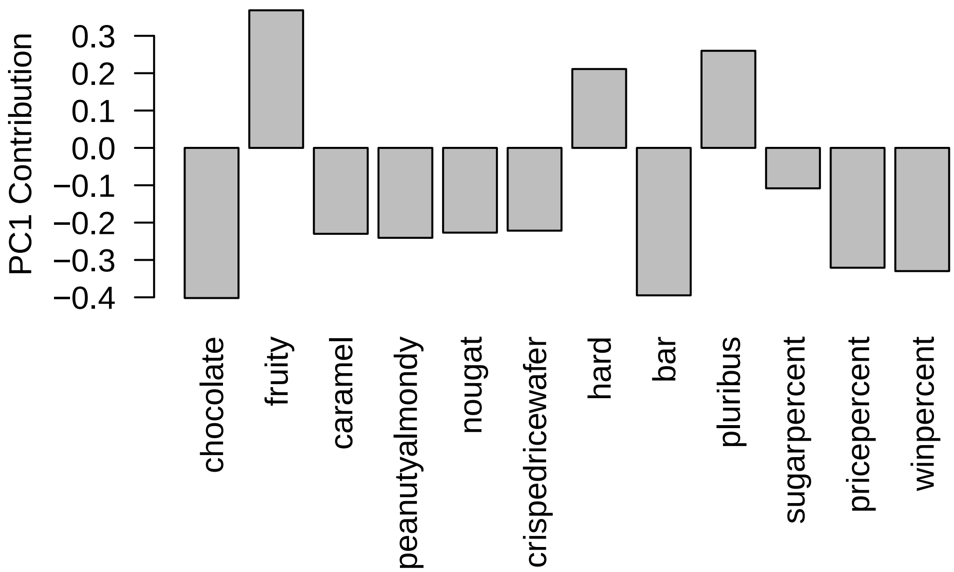

```
pc <- as.data.frame(pca$x)

ggplot(pc) +
  aes(x=PC1, y=PC2, label=rownames(pc)) +
  geom_point(col=my_cols) +
  geom_text_repel(max.overlaps = 5)
```

```
Warning: ggrepel: 71 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



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

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



Fruity. PC1 is strongly correlated with fruity variable