Class 10: Halloween Mini-Project

Rosa Chavez (PID: A16487039)

```
options(repos = c(CRAN="http://cran.us.r-project.org"))
```

```
candy_file <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-r
```

```
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?

```
n_candy_types <- nrow(candy)
n_candy_types
```

```
[1] 85
```

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

```
n_fruity_candies <- sum(candy$fruity == 1)
n_fruity_candies
```

[1] 38

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

Hershey's Special Dark

```
winpercent_fav <- candy["Hershey's Special Dark", "winpercent"]

winpercent_fav
```

[1] 59.23612

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

```
winpercent_kit_kat <- candy["Kit Kat", "winpercent"]
winpercent_kit_kat
```

[1] 76.7686

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

```
winpercent_tootsie_roll <- candy["Tootsie Roll Snack Bars", "winpercent"]
winpercent_tootsie_roll
```

[1] 49.6535

```
install.packages("skimr")
```

Installing package into 'C:/Users/ITSloaner.DESKTOP-I89K3M9/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)

package 'skimr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\ITSloaner.DESKTOP-I89K3M9\AppData\Local\Temp\RtmpQjtUVP\downloaded_packages

```r
library("skimr")

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, the last one winpercent does not have numbers 0 to 1 but rather 0 to 100.

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

1 represents candies that contain chocolates while 0 represents candies that do not contain chocolate.

Q8. Plot a histogram of winpercent values

```r
hist(candy$winpercent, main="Histogram of Winpercent", xlab="Win Percent")
```

Histogram of Winpercent
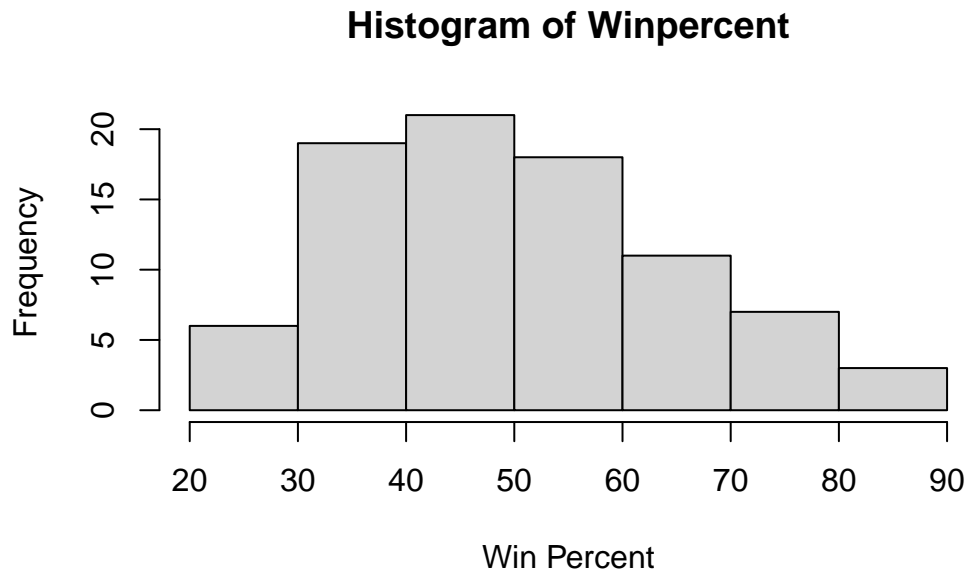
Q9. Is the distribution of winpercent values symmetrical?

The distribution is slightly skewed to the right.

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

Center is above 50%.

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

```r
avg_chocolate <- mean(candy$winpercent[candy$chocolate == 1])
avg_fruity <- mean(candy$winpercent[candy$fruity == 1])
avg_chocolate
```

```
[1] 60.92153
```

```r
avg_fruity
```

```
[1] 44.11974
```

chocolate is higher ranked than fruity candy.

Q12. Is this difference statistically significant?

```r
t.test(candy$winpercent[candy$chocolate == 1], candy$winpercent[candy$fruity == 1])
```

```
    Welch Two Sample t-test

data:  candy$winpercent[candy$chocolate == 1] and candy$winpercent[candy$fruity == 1]
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?

```r
least_liked <- head(candy[order(candy$winpercent), ], n = 5)
least_liked
```

```
                  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?

```r
most_liked <- head(candy[order(candy$winpercent, decreasing = TRUE), ], n = 5)
most_liked
```

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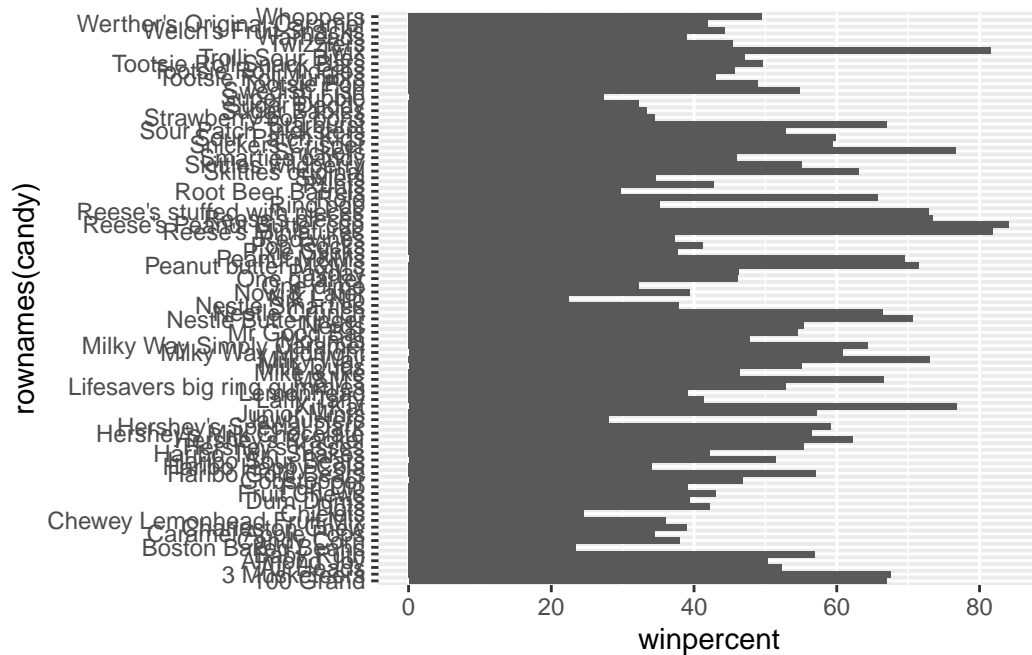
Q15. Make a first barplot of candy ranking based on winpercent values.

```r
library(ggplot2)
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_bar(stat = "identity")
```

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

```r
library(ggplot2)
ggplot(candy, aes(x = reorder(rownames(candy), -winpercent), y = winpercent)) +
geom_bar(stat = "identity") +
coord_flip() +
theme_minimal()
```
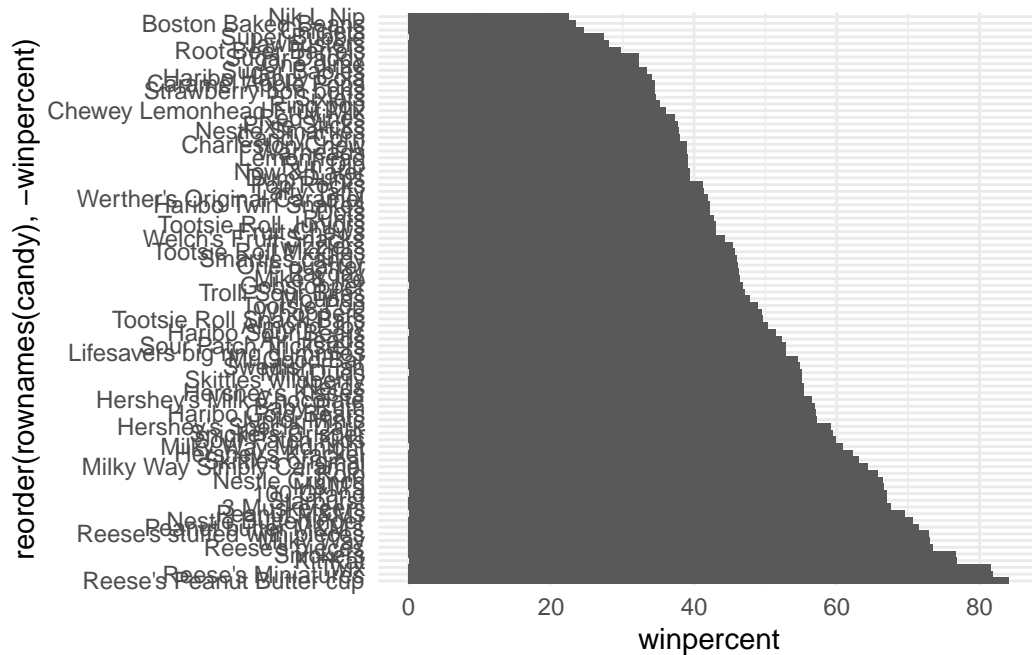
Q17. What is the worst ranked chocolate candy?

```
worst_chocolate <- candy[candy$chocolate == 1, ] [which.min(candy[candy$chocolate == 1, "w
worst_chocolate
```

	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. What is the best ranked fruity candy?

```
best_fruity <- candy[candy$fruity == 1, ] [which.max(candy[candy$fruity == 1, "winpercent"
best_fruity
```

	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

```
install.packages("ggrepel")
```

Installing package into 'C:/Users/ITSloaner.DESKTOP-I89K3M9/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)

package 'ggrepel' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\ITSloaner.DESKTOP-I89K3M9\AppData\Local\Temp\RtmpQjtUVP\downloaded_packages

```
library(ggrepel)
```

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?

```
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, 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
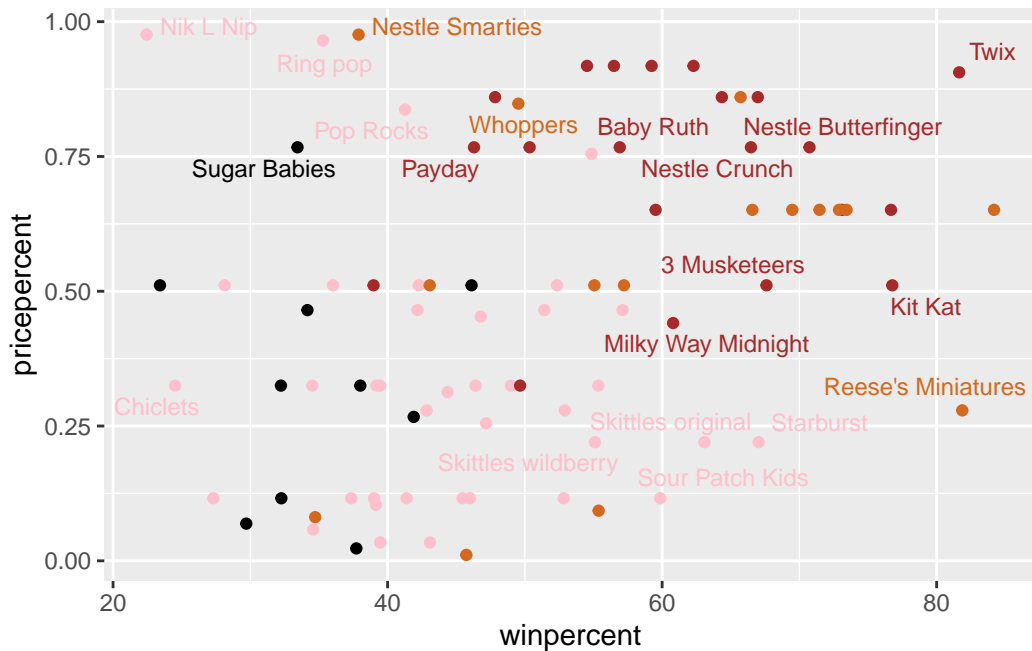increasing max.overlaps
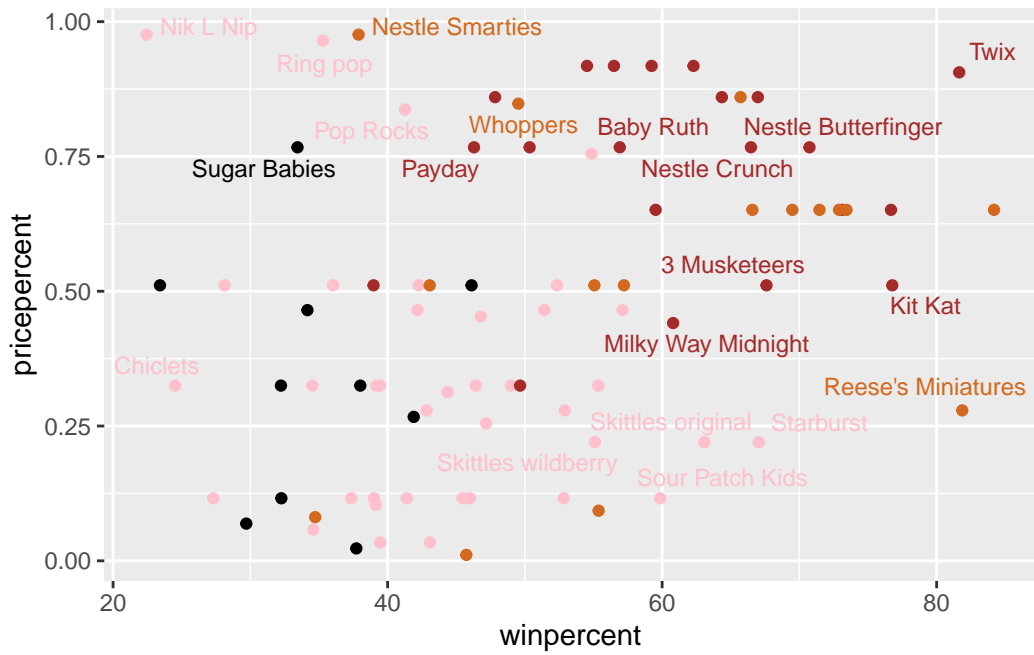
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)
```
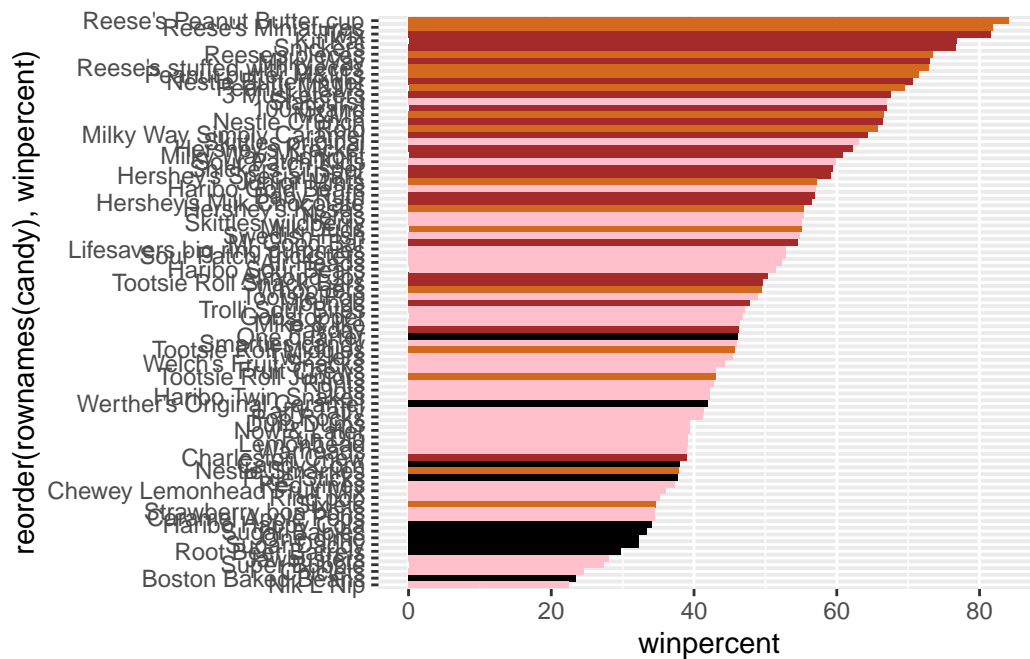
```
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 5)
```

Warning: ggrepel: 65 unlabeled data points (too many overlaps). Consider
increasing max.overlaps

Reese'e Minatures

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

```
ord <- order(candy$pricepercent, decreasing = TRUE)
expensive_candies <- head(candy[ord, ], n = 5)
print(expensive_candies)
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Nik L Nip	0	1	0	0	0
Nestle Smarties	1	0	0	0	0
Ring pop	0	1	0	0	0
Hershey's Krackel	1	0	0	0	0
Hershey's Milk Chocolate	1	0	0	0	0

	crispedricewafer	hard	bar	pluribus	sugarpercent
Nik L Nip	0	0	0	1	0.197
Nestle Smarties	0	0	0	1	0.267
Ring pop	0	1	0	0	0.732
Hershey's Krackel	1	0	1	0	0.430
Hershey's Milk Chocolate	0	0	1	0	0.430

                         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
```

```r
install.packages("corrplot")
```

```
Installing package into 'C:/Users/ITSloaner.DESKTOP-I89K3M9/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)

package 'corrplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\ITSloaner.DESKTOP-I89K3M9\AppData\Local\Temp\RtmpQjtUVP\downloaded_packages
```
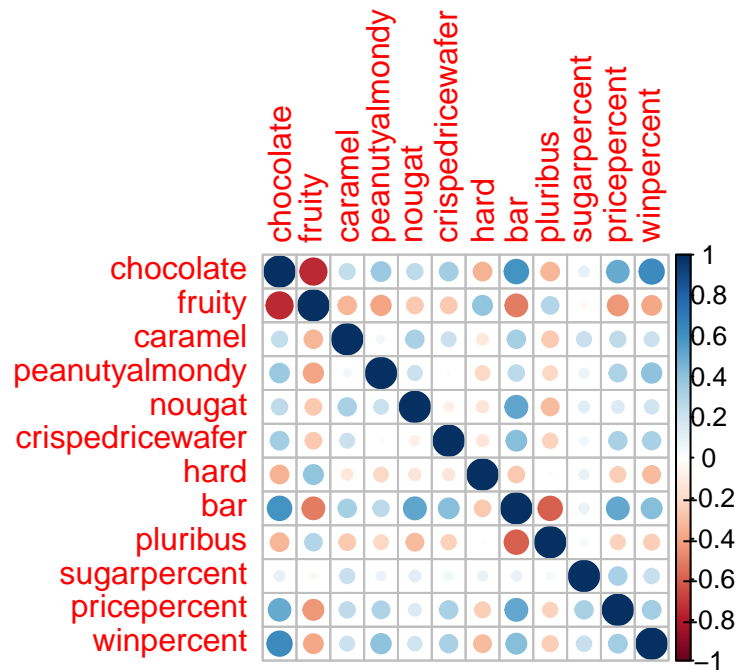
```r
library(corrplot)
```

```
corrplot 0.92 loaded
```

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

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

Chocolate and sugarpercent.

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

Chocolate and nougat.

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

```
pca <- prcomp(candy[, -1], scale. = TRUE)
summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     1.9200 1.1143 1.1085 1.0751 0.95010 0.81815 0.81352
Proportion of Variance 0.3351 0.1129 0.1117 0.1051 0.08206 0.06085 0.06016
Cumulative Proportion  0.3351 0.4480 0.5597 0.6648 0.74685 0.80770 0.86787
                          PC8     PC9    PC10    PC11
Standard deviation     0.68950 0.64410 0.60875 0.43887
Proportion of Variance 0.04322 0.03772 0.03369 0.01751
Cumulative Proportion  0.91109 0.94880 0.98249 1.00000
```

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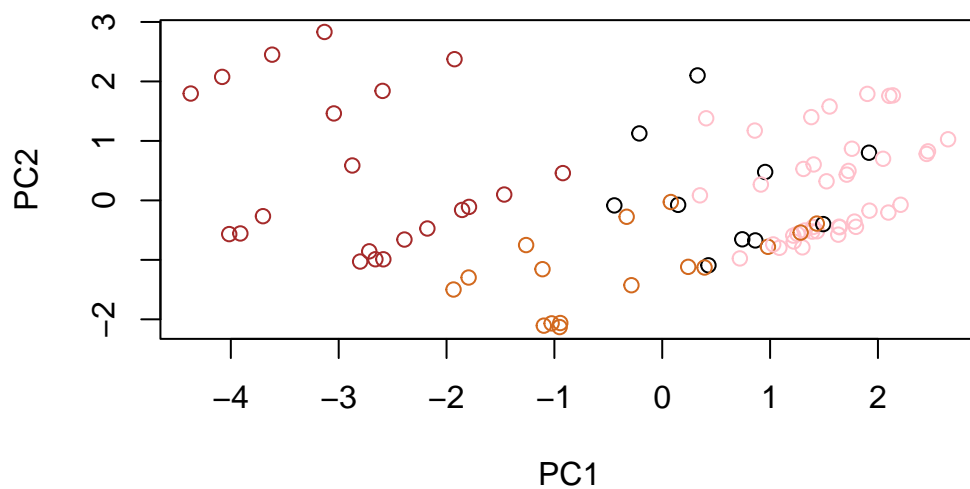


```
# Adding color and labels to PCA plot
ggplot(as.data.frame(pca$x[, 1:2]), aes(x = PC1, y = PC2, color = my_cols, label = rowname
geom_point() +
geom_text_repel()
```

Warning: ggrepel: 54 unlabeled data points (too many overlaps). Consider
increasing max.overlaps