Class 11 - Candy

Emmanuel Robles

```
candy <- read.csv("candy-data.csv", 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
```

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

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

```
[1] 38
```

What are these fruity candies?

```
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"
```

```
rownames( candy[ candy$chocolate == 1, ])
```

```
 [1] "100 Grand"                    "3 Musketeers"
 [3] "Almond Joy"                   "Baby Ruth"
 [5] "Charleston Chew"              "Hershey's Kisses"
 [7] "Hershey's Krackel"            "Hershey's Milk Chocolate"
 [9] "Hershey's Special Dark"       "Junior Mints"
[11] "Kit Kat"                      "Peanut butter M&M's"
[13] "M&M's"                        "Milk Duds"
[15] "Milky Way"                    "Milky Way Midnight"
[17] "Milky Way Simply Caramel"     "Mounds"
[19] "Mr Good Bar"                  "Nestle Butterfinger"
[21] "Nestle Crunch"                "Peanut M&Ms"
[23] "Reese's Miniatures"           "Reese's Peanut Butter cup"
[25] "Reese's pieces"               "Reese's stuffed with pieces"
[27] "Rolo"                         "Sixlets"
[29] "Nestle Smarties"              "Snickers"
[31] "Snickers Crisper"             "Tootsie Pop"
```

```
[33] "Tootsie Roll Juniors"        "Tootsie Roll Midgies"
[35] "Tootsie Roll Snack Bars"     "Twix"
[37] "Whoppers"
```

How often does my favorite candy win?

Q3. What is your favorite candy in the dataset and what is it's winpercent value? Mine is higher than professors. Superior candy, clearly.

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

```
[1] 81.64291
```

```
candy["Reese's Peanut Butter cup", "winpercent"]
```

```
[1] 84.18029
```

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

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

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

Table 1: Data summary

Name	candy
Number of rows	85
Number of columns	12

Table 1: Data summary

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	

```
skimr::skim(candy)
```

Table 3: 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 `winpercent` column is on a 0:100 scale and all others appear to be on a 0:1 scale.

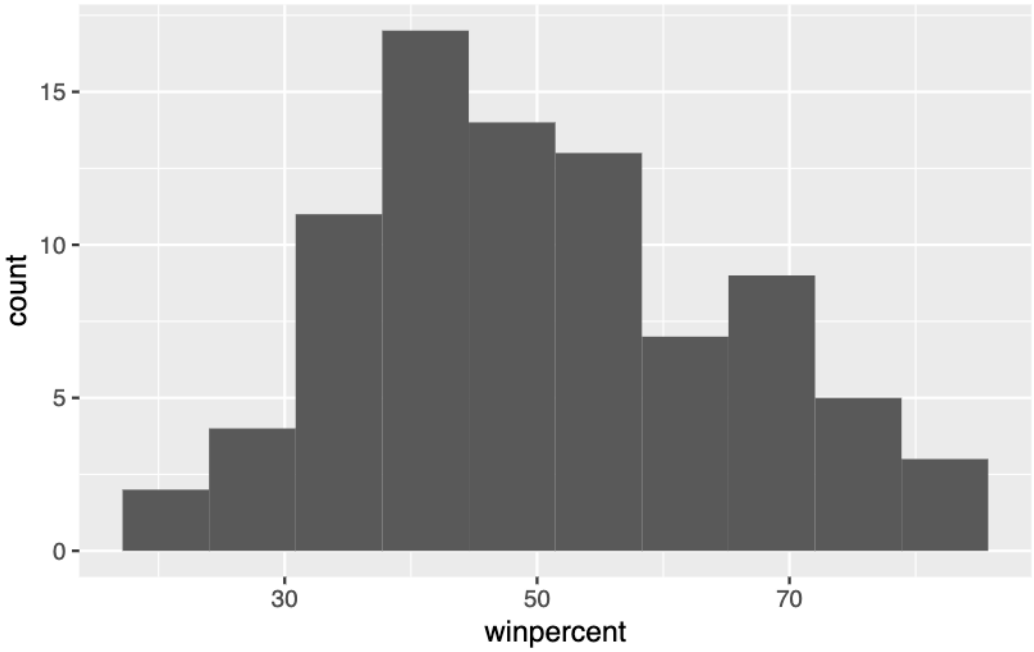
Q7. What do you think a zero and one represent for the candy$chocolate column?

A zero here means a candy is not classified as containing chocolate.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)

ggplot(candy, aes(winpercent)) +
  geom_histogram(bins=10)
```



Q9. Is the distribution of winpercent values symmetrical?

No

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

Below 50% with a mean of

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

```
[1] 50.31676
```

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

Fom the averages it looks like people prefer chocolate candy over fruity.

```
fruit.mean <- mean(candy[candy$fruity == 1, ]$winpercent)
choco.mean <- mean(candy[candy$chocolate == 1, ]$winpercent)
fruit.mean
```

```
[1] 44.11974
```

```
  choco.mean
```

```
[1] 60.92153
```

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

```
    Welch Two Sample t-test

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

Overall Candy Rankings

There is a base function in R called `sort()` for, guess what, sorting vector inputs. > Q13. What are the five least liked candy types in this set?

```
  candy.rank <- sort(candy$winpercent, decreasing= TRUE)
  candy.rank
```

```
 [1] 84.18029 81.86626 81.64291 76.76860 76.67378 73.43499 73.09956 72.88790
 [9] 71.46505 70.73564 69.48379 67.60294 67.03763 66.97173 66.57458 66.47068
[17] 65.71629 64.35334 63.08514 62.28448 60.80070 59.86400 59.52925 59.23612
[25] 57.21925 57.11974 56.91455 56.49050 55.37545 55.35405 55.10370 55.06407
[33] 54.86111 54.52645 52.91139 52.82595 52.34146 51.41243 50.34755 49.65350
[41] 49.52411 48.98265 47.82975 47.17323 46.78335 46.41172 46.29660 46.11650
[49] 45.99583 45.73675 45.46628 44.37552 43.08892 43.06890 42.84914 42.27208
[57] 42.17877 41.90431 41.38956 41.26551 39.46056 39.44680 39.18550 39.14106
[65] 39.01190 38.97504 38.01096 37.88719 37.72234 37.34852 36.01763 35.29076
[73] 34.72200 34.57899 34.51768 34.15896 33.43755 32.26109 32.23100 29.70369
[81] 28.12744 27.30386 24.52499 23.41782 22.44534
```

The buddy funciton to `sort()` this is often more useful is called `order()`.

I can order by `winpercent`.

```
ord <- order(candy$winpercent)
ord
```

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

```
head(candy[ord,])
```

```
                  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
Root Beer Barrels         0      0       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
Root Beer Barrels                0    1   0        1        0.732        0.069
                  winpercent
Nik L Nip           22.44534
Boston Baked Beans  23.41782
Chiclets            24.52499
Super Bubble        27.30386
Jawbusters          28.12744
Root Beer Barrels   29.70369
```

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

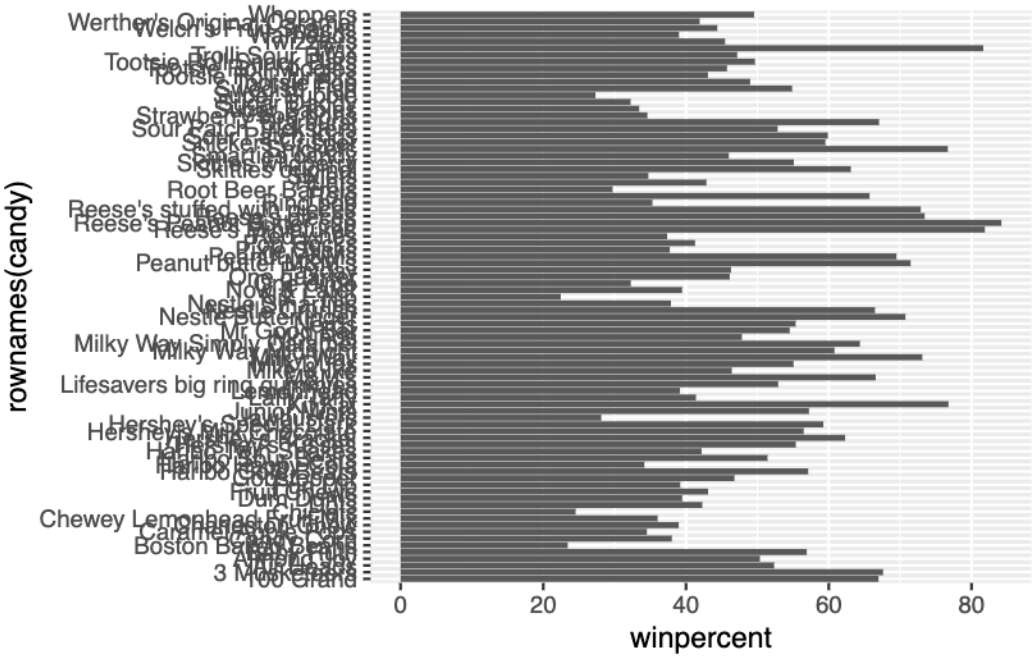
```
ord1  <- order(candy$winpercent, decreasing= TRUE)
head(candy[ord1,])
```

```
                         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
Reese's pieces                    1      0       0              1      0
                         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
Reese's pieces                           0    0   0        1        0.406
                         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
Reese's pieces                   0.651   73.43499
```

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

```
library(ggplot2)

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

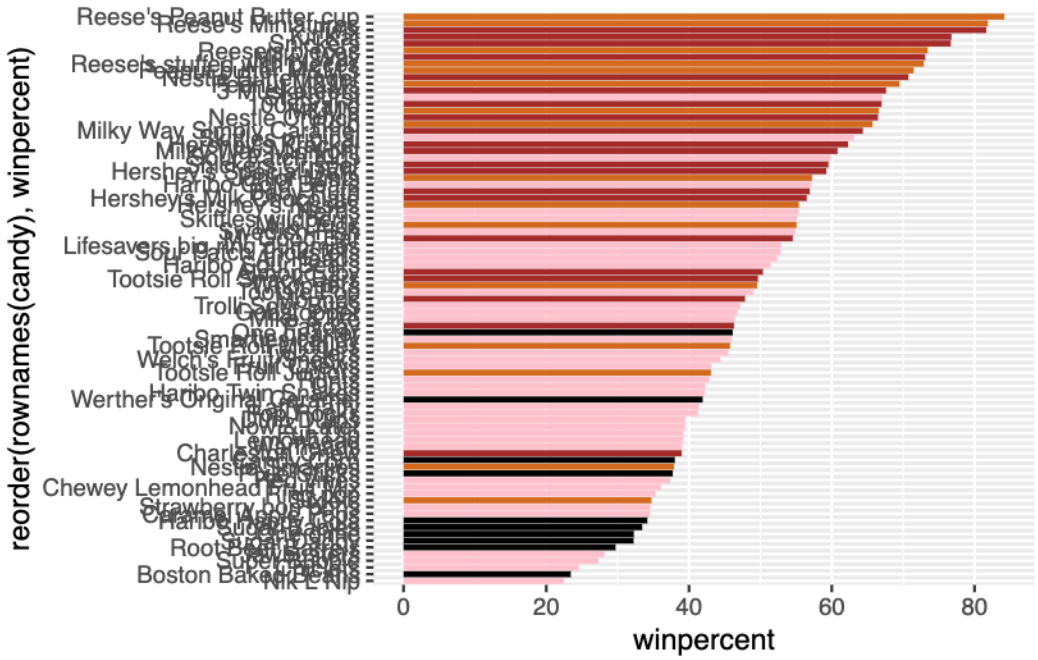


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
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"
```

```
library(ggplot2)

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



Q17. What is the worst ranked chocolate candy?

Sixlets

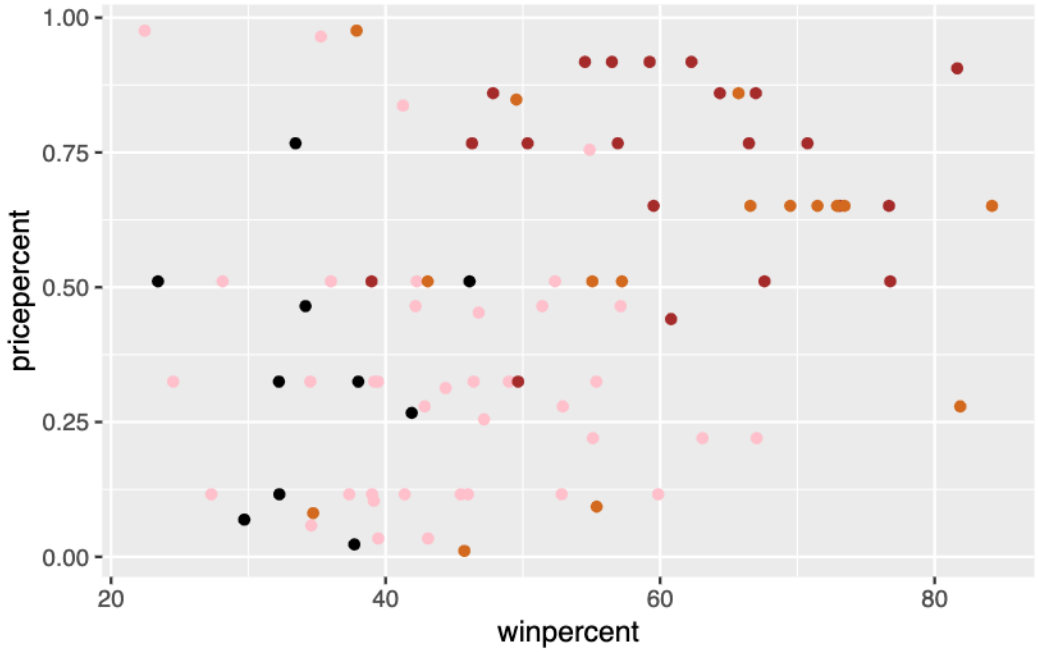
Q18. What is the best ranked fruity candy?

Starburst

Taking a look at pricepoint

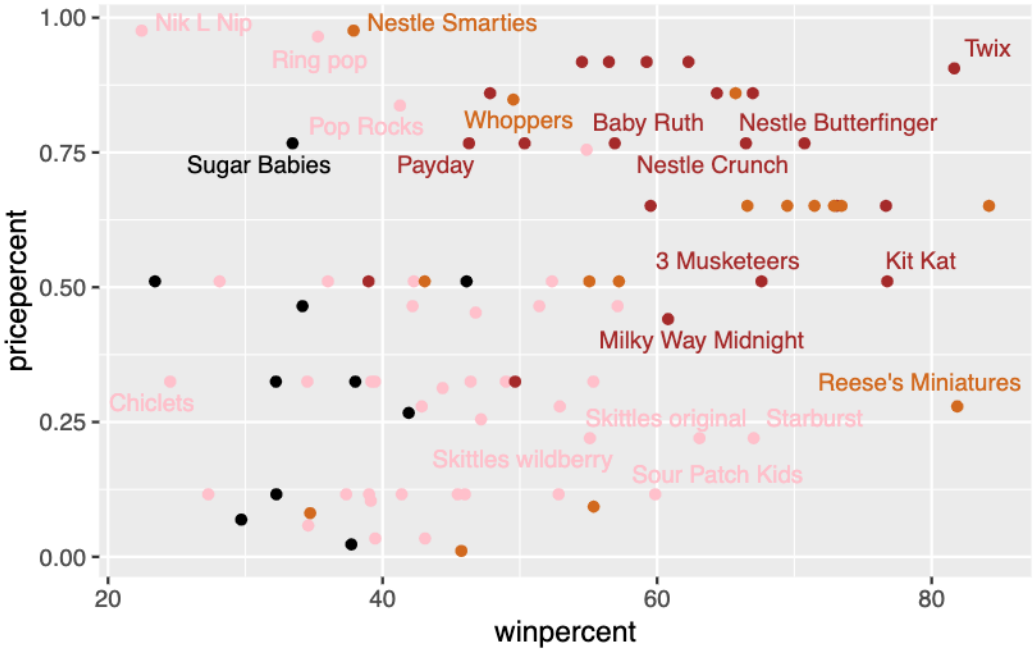
What is the best candy for the least money?

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




```
library(ggrepel)
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?

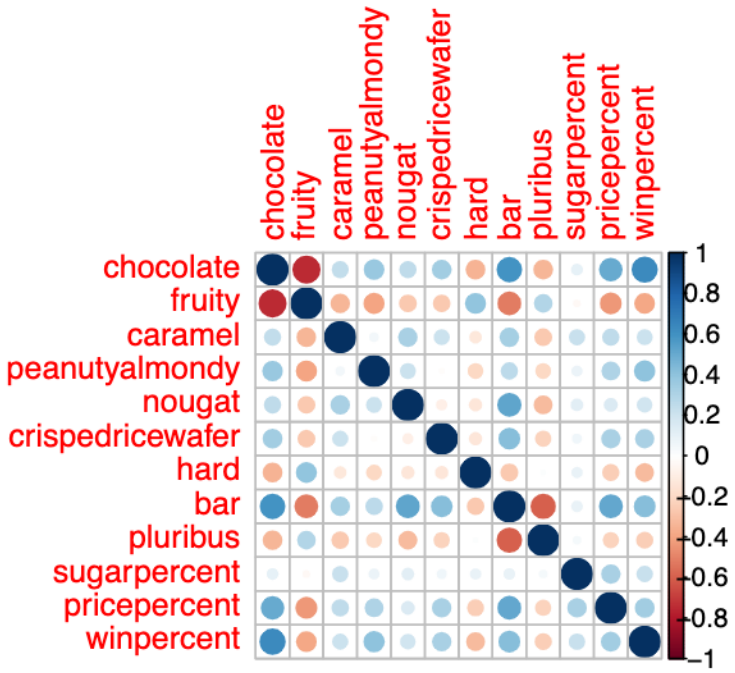
5 Exploring the correlation structure

Pearson correlation goes between -1 and +1 with zero indicating no correlation and values close to one being very highly (an) correlated

```
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)?

Chocolate and fruit.

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

Chocolate and winpercent or chocolate and bar.

6 Principal Component Analysis.

The base R function for PCA is called `prcomp()` and we can set "scale= TRUE/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
```

The main result of PCA - i.e. the new PC plot (projection of candy in our new PC axis) is contained in `pca$x`

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

ggplot(pc)+
  aes(PC1, PC2)+
  geom_point(col=my_cols)
```