Class 10 Halloween

Yu-Chia Huang (PID: A59026739)

##Importing candy data

```r
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?

```r
nrow(candy)
```

```
[1] 85
```

There are 85.

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

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

```
[1] 38
```

```r
candy[as.logical(candy$fruity),]
```

```
                            chocolate fruity caramel peanutyalmondy nougat
Air Heads                           0      1       0              0      0
Caramel Apple Pops                  0      1       1              0      0
Chewey Lemonhead Fruit Mix          0      1       0              0      0
Chiclets                            0      1       0              0      0
Dots                                0      1       0              0      0
Dum Dums                            0      1       0              0      0
Fruit Chews                         0      1       0              0      0
Fun Dip                             0      1       0              0      0
Gobstopper                          0      1       0              0      0
Haribo Gold Bears                   0      1       0              0      0
Haribo Sour Bears                   0      1       0              0      0
Haribo Twin Snakes                  0      1       0              0      0
Jawbusters                          0      1       0              0      0
Laffy Taffy                         0      1       0              0      0
Lemonhead                           0      1       0              0      0
Lifesavers big ring gummies         0      1       0              0      0
Mike & Ike                          0      1       0              0      0
Nerds                               0      1       0              0      0
Nik L Nip                           0      1       0              0      0
Now & Later                         0      1       0              0      0
Pop Rocks                           0      1       0              0      0
Red vines                           0      1       0              0      0
Ring pop                            0      1       0              0      0
Runts                               0      1       0              0      0
Skittles original                   0      1       0              0      0
Skittles wildberry                  0      1       0              0      0
Smarties candy                      0      1       0              0      0
Sour Patch Kids                     0      1       0              0      0
Sour Patch Tricksters               0      1       0              0      0
Starburst                           0      1       0              0      0
Strawberry bon bons                 0      1       0              0      0
Super Bubble                        0      1       0              0      0
Swedish Fish                        0      1       0              0      0
```

```
Tootsie Pop                         1      1       0              0      0
Trolli Sour Bites                   0      1       0              0      0
Twizzlers                           0      1       0              0      0
Warheads                            0      1       0              0      0
Welch's Fruit Snacks                0      1       0              0      0
                            crispedricewafer hard bar pluribus sugarpercent
Air Heads                                  0    0   0        0        0.906
Caramel Apple Pops                         0    0   0        0        0.604
Chewey Lemonhead Fruit Mix                 0    0   0        1        0.732
Chiclets                                   0    0   0        1        0.046
Dots                                       0    0   0        1        0.732
Dum Dums                                   0    1   0        0        0.732
Fruit Chews                                0    0   0        1        0.127
Fun Dip                                    0    1   0        0        0.732
Gobstopper                                 0    1   0        1        0.906
Haribo Gold Bears                          0    0   0        1        0.465
Haribo Sour Bears                          0    0   0        1        0.465
Haribo Twin Snakes                         0    0   0        1        0.465
Jawbusters                                 0    1   0        1        0.093
Laffy Taffy                                0    0   0        0        0.220
Lemonhead                                  0    1   0        0        0.046
Lifesavers big ring gummies                0    0   0        0        0.267
Mike & Ike                                 0    0   0        1        0.872
Nerds                                      0    1   0        1        0.848
Nik L Nip                                  0    0   0        1        0.197
Now & Later                                0    0   0        1        0.220
Pop Rocks                                  0    1   0        1        0.604
Red vines                                  0    0   0        1        0.581
Ring pop                                   0    1   0        0        0.732
Runts                                      0    1   0        1        0.872
Skittles original                          0    0   0        1        0.941
Skittles wildberry                         0    0   0        1        0.941
Smarties candy                             0    1   0        1        0.267
Sour Patch Kids                            0    0   0        1        0.069
Sour Patch Tricksters                      0    0   0        1        0.069
Starburst                                  0    0   0        1        0.151
Strawberry bon bons                        0    1   0        1        0.569
Super Bubble                               0    0   0        0        0.162
Swedish Fish                               0    0   0        1        0.604
Tootsie Pop                                0    1   0        0        0.604
Trolli Sour Bites                          0    0   0        1        0.313
Twizzlers                                  0    0   0        0        0.220
Warheads                                   0    1   0        0        0.093
```

```
Welch's Fruit Snacks                       0    0   0        1        0.313
                            pricepercent winpercent
Air Heads                          0.511   52.34146
Caramel Apple Pops                 0.325   34.51768
Chewey Lemonhead Fruit Mix         0.511   36.01763
Chiclets                           0.325   24.52499
Dots                               0.511   42.27208
Dum Dums                           0.034   39.46056
Fruit Chews                        0.034   43.08892
Fun Dip                            0.325   39.18550
Gobstopper                         0.453   46.78335
Haribo Gold Bears                  0.465   57.11974
Haribo Sour Bears                  0.465   51.41243
Haribo Twin Snakes                 0.465   42.17877
Jawbusters                         0.511   28.12744
Laffy Taffy                        0.116   41.38956
Lemonhead                          0.104   39.14106
Lifesavers big ring gummies        0.279   52.91139
Mike & Ike                         0.325   46.41172
Nerds                              0.325   55.35405
Nik L Nip                          0.976   22.44534
Now & Later                        0.325   39.44680
Pop Rocks                          0.837   41.26551
Red vines                          0.116   37.34852
Ring pop                           0.965   35.29076
Runts                              0.279   42.84914
Skittles original                  0.220   63.08514
Skittles wildberry                 0.220   55.10370
Smarties candy                     0.116   45.99583
Sour Patch Kids                    0.116   59.86400
Sour Patch Tricksters              0.116   52.82595
Starburst                          0.220   67.03763
Strawberry bon bons                0.058   34.57899
Super Bubble                       0.116   27.30386
Swedish Fish                       0.755   54.86111
Tootsie Pop                        0.325   48.98265
Trolli Sour Bites                  0.255   47.17323
Twizzlers                          0.116   45.46628
Warheads                           0.116   39.01190
Welch's Fruit Snacks               0.313   44.37552
```

##2. What is your favorate candy?

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

```r
candy["Hershey's Kisses",]$winpercent
```

```
[1] 55.37545
```

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

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

##Side-note: the skimr::skim() function

#Use :: then we can do the same thing (install the package and call it). #install.packages("skimr") #library("skimr")

```r
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?

Sure is, Winpercent.

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

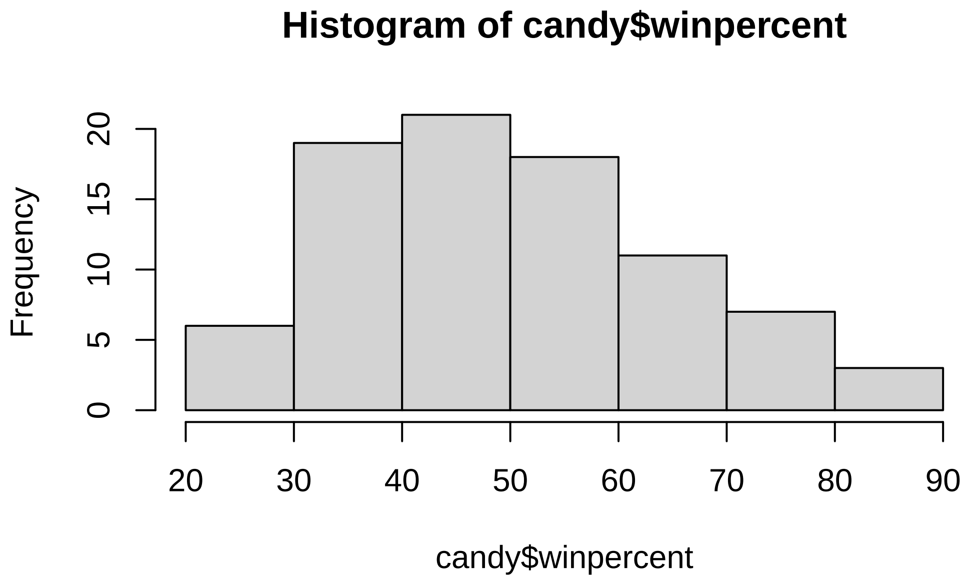
```r
candy$chocolate
```

```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

It represents whether each candy is chocolate (1) or not chocolate(0).

> Q8. Plot a histogram of winpercent values

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



> Q9. Is the distribution of winpercent values symmetrical?

No. It's not symmetrical.

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

```
choc.ind <- as.logical(candy$chocolate)
fruit.ind <- as.logical(candy$fruity)

choc.win <- candy[choc.ind,]$winpercent
fruit.win  <- candy[fruit.ind,]$winpercent

mean(candy$winpercent)
```

```
[1] 50.31676
```

It's above 50%.

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

```
mean(choc.win)
```

```
[1] 60.92153
```

```r
mean(fruit.win)
```

```
[1] 44.11974
```

Chocolate is higher ranked than fruit candy.

> Q12. Is this difference statistically significant?

```r
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 
```

Yes. The p-value is 2.871e-08. We can reject the null hypothesis.

##3. Overall Candy Rankings >Q13. What are the five least liked candy types in this set?

```r
x <- c(5,2,3,6)
sort(x)
```

```
[1] 2 3 5 6
```

```r
sort(x, decreasing = T)
```

```
[1] 6 5 3 2
```

```r
x
```

```
[1] 5 2 3 6
```

```
order(x)
```

```
[1] 2 3 1 4
```

```
x[order(x)]
```

```
[1] 2 3 5 6
```

```
y <- c("D", "A", "E")
order(y)
```

```
[1] 2 1 3
```

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

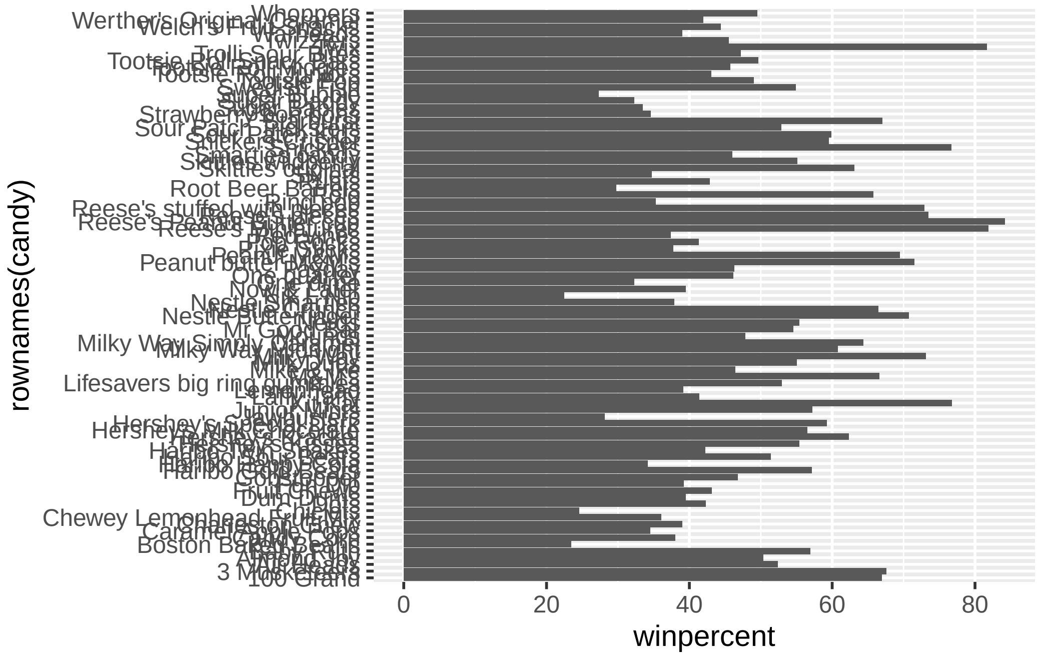
Q14. What are the top 5 all time favorite candy types out of this set?

```
inds <- order(candy$winpercent, decreasing = T)
head(candy[inds,], 5)
```

```
                         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
```

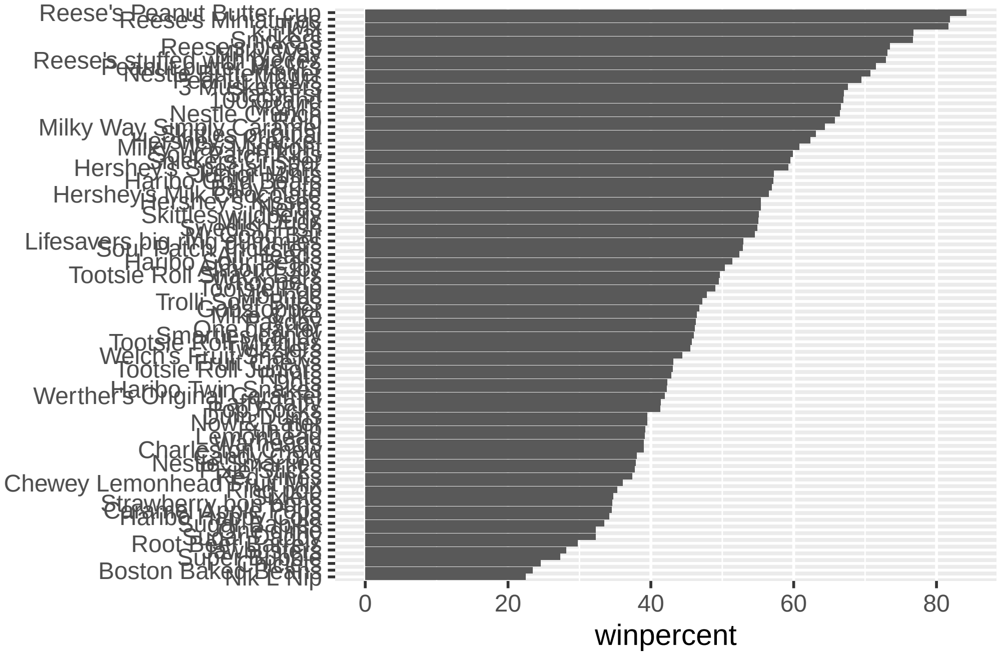
Q15. Make a first barplot of candy ranking based on winpercent values.

```
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?

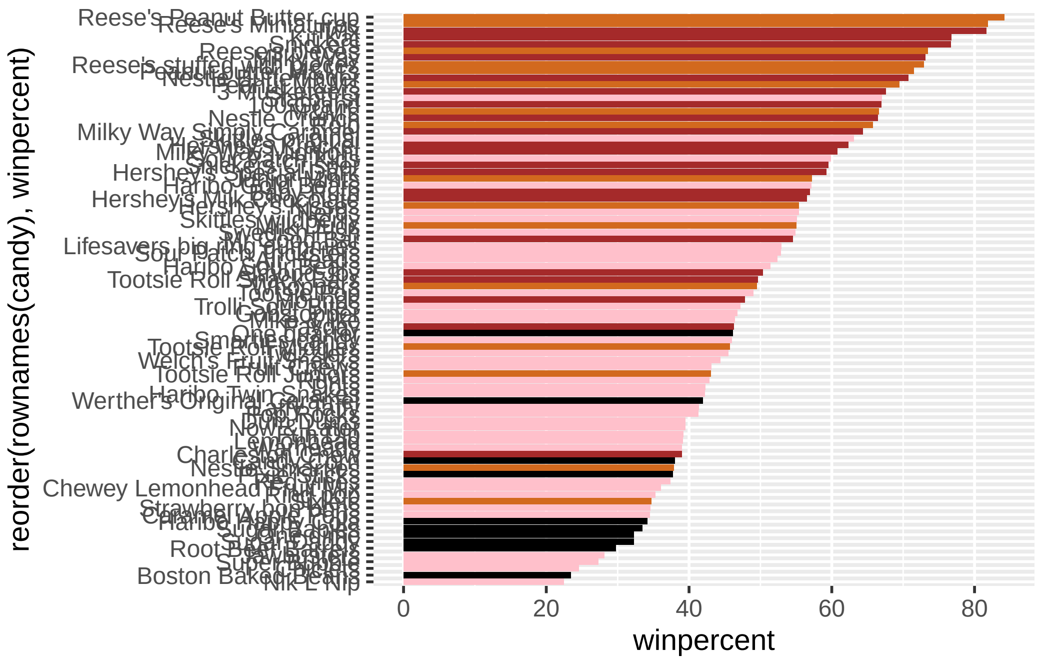
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col()
```



##Time to add some useful color Let's color up these candy up by some scheme.

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

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



Now, for the first time, using this plot we can answer questions like:

> Q17. What is the worst ranked chocolate candy?

Sixlets is the worst ranked chocolate candy.

> Q18. What is the best ranked fruity candy?

Starburst is the best ranked fruity candy.

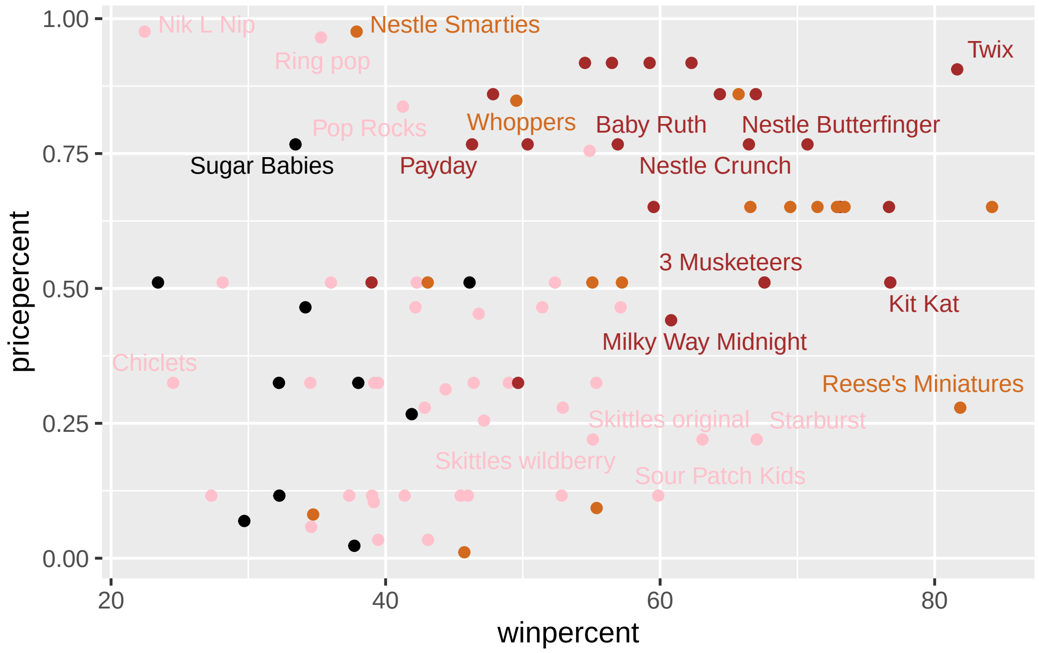
##4. Taking a look at pricepercent

> 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?

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

```
library("ggrepel")
library("ggplot2")
# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=mycols) +
    geom_text_repel(col=mycols, size=3.3, max.overlaps = 5)
```

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



Reese's Miniatures.

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

```
                     pricepercent winpercent
Tootsie Roll Midgies        0.011   45.73675
Pixie Sticks                0.023   37.72234
Dum Dums                    0.034   39.46056
Fruit Chews                 0.034   43.08892
Strawberry bon bons         0.058   34.57899
```

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

Nik L Ni, Ring pop, Nestle Smarties, Pop Rocks, Sugar Babies.

```r
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,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
```

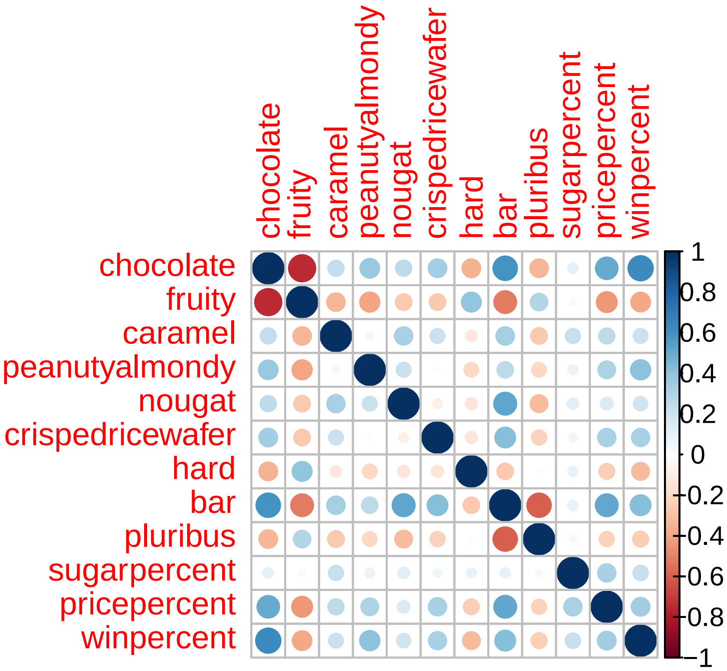
Optional >Q21. Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called “dot chat” or “lollipop” chart by swapping geom_col() for geom_point() + geom_segment().

##5 Exploring the correlation structure # Correlation structure

```r
library("corrplot")
```

```
corrplot 0.92 loaded
```

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



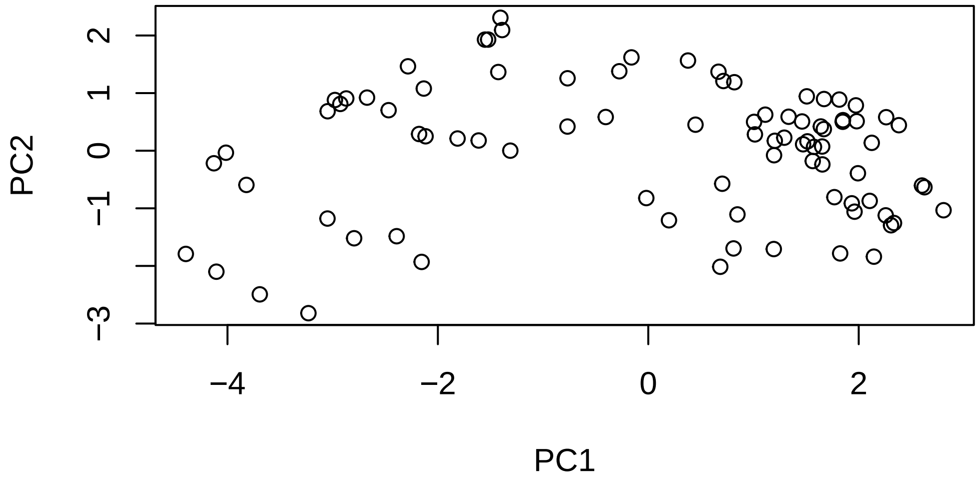
> Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?
>
> Q23. Similarly, what two variables are most positively correlated?

##6. Principal Component Analysis

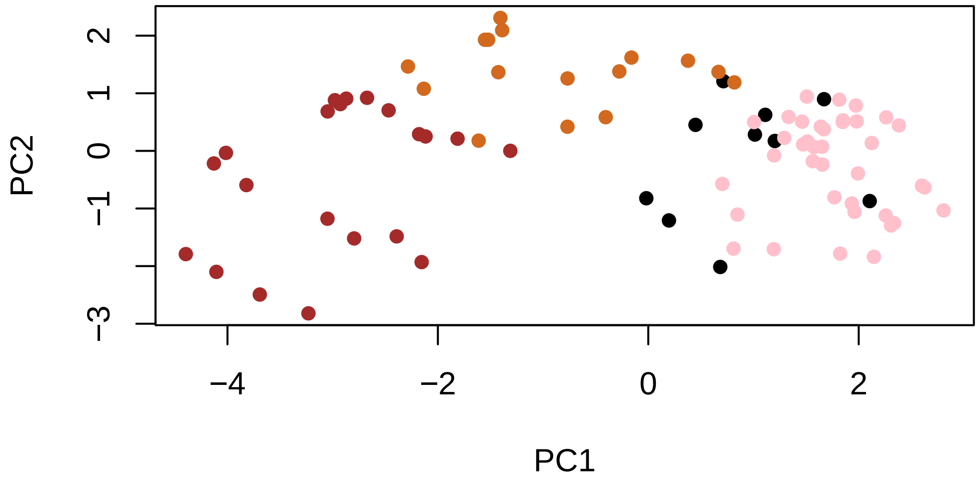
```
pca <- prcomp(candy,scale=TRUE)
summary(pca)
```

```
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
```

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



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


PC2
PC1