# Class 9 - Halloween - BIMM 143

Abzael Verduzco

##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?

```
dim(candy)
```

```
[1] 85 12
```

There are 85 different candy types.

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

```
sum(candy[,2])
```

```
[1] 38
```

```
#or
sum(candy$fruity)
```

```
[1] 38
```

There are 38 fruity candies.

##2. What is your favorate candy?

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

```
[1] 81.64291
```

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

```
#My favorite candy is Reeses

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

```
[1] 84.18029
```

84% is its win percent value.

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

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

```
[1] 76.7686
```

It has a win percentage of roughly 77% percent.

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

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

```
[1] 49.6535
```

It has a win percentage of roughly 50 percent.

```
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?

The variable hist in a different scale than the rest of columns in the dataset, it is not numeric.

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

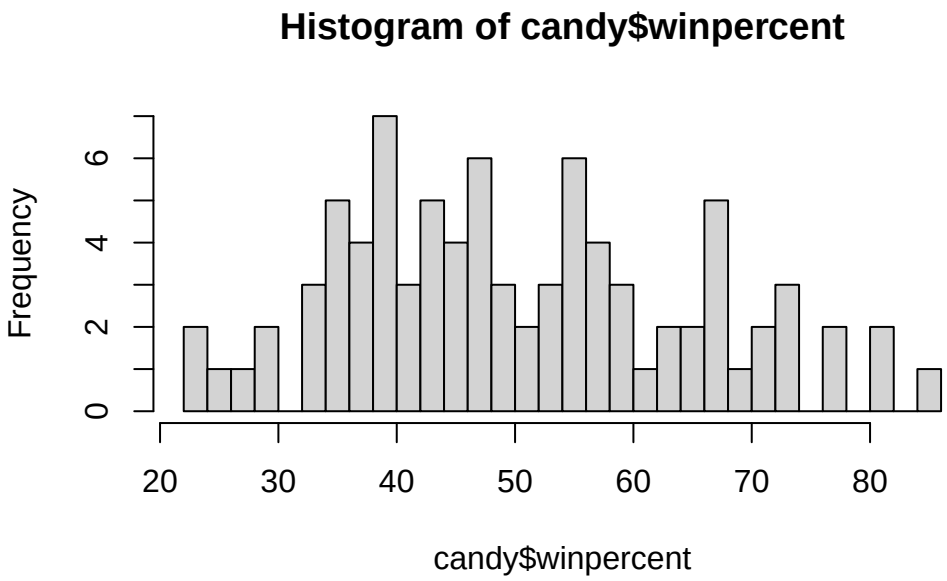
```
candy$chocolate
```

```
 [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
```

1 represent that chocolate is present, and 0 represent that it is not present.

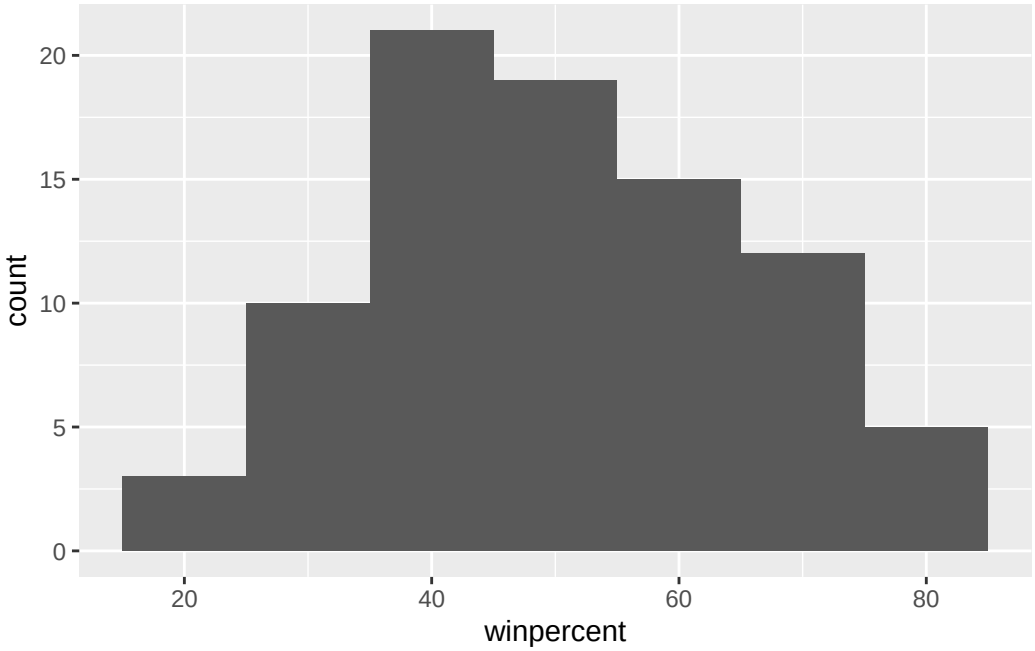
Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent, breaks=30)
```




```
library(ggplot2)
```

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




Q9. Is the distribution of winpercent values symmetrical?

The distribution of win values is not symmetrical.

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

The center of distribution is below 50%.

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

```
#Chocolate
choc.inds <- as.logical(candy$chocolate)
choc.win <- candy[choc.inds, 'winpercent']
choc.win
```

```
 [1] 66.97173 67.60294 50.34755 56.91455 38.97504 55.37545 62.28448 56.49050
 [9] 59.23612 57.21925 76.76860 71.46505 66.57458 55.06407 73.09956 60.80070
[17] 64.35334 47.82975 54.52645 70.73564 66.47068 69.48379 81.86626 84.18029
[25] 73.43499 72.88790 65.71629 34.72200 37.88719 76.67378 59.52925 48.98265
[33] 43.06890 45.73675 49.65350 81.64291 49.52411
```

```
#Fruity
fruity.inds <- as.logical(candy$fruity)
```

```
  fruity.win <- candy[fruity.inds, 'winpercent']
  fruity.win
```

```
 [1] 52.34146 34.51768 36.01763 24.52499 42.27208 39.46056 43.08892 39.18550
 [9] 46.78335 57.11974 51.41243 42.17877 28.12744 41.38956 39.14106 52.91139
[17] 46.41172 55.35405 22.44534 39.44680 41.26551 37.34852 35.29076 42.84914
[25] 63.08514 55.10370 45.99583 59.86400 52.82595 67.03763 34.57899 27.30386
[33] 54.86111 48.98265 47.17323 45.46628 39.01190 44.37552
```

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

```
[1] 60.92153
```

```
  mean(fruity.win)
```

```
[1] 44.11974
```

On average the chocolate candy is higher than the fruit candy.

Q12. Is this difference statistically significant?

```
  t.test(choc.win,fruity.win)
```

```
    Welch Two Sample t-test

data:  choc.win and fruity.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
```

Then it is statistical significance.

## 3. Overall Candy Rankings

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

```
head(candy[order(candy$winpercent),], 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
```

Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble and Jawbusters are the least liked candies in this dataset.

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

```
head(candy[order(candy$winpercent, decreasing=TRUE),], n=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
```

The top 5 all time favorite candy types out of this set are Snickers, Kit Kat, Twix, Reese's Miniatures, and Reese's Peanut Butter Cup.

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
candy %>% arrange(winpercent) %>% head(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
```

```
candy %>% arrange(winpercent) %>% tail(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
Reese's Miniatures                1      0       0              1      0
Reese's 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
Reese's Miniatures                        0    0   0        0        0.034
Reese's 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
Reese's Miniatures               0.279   81.86626
Reese's Peanut Butter cup        0.651   84.18029
```

In my opinion I like the order method more as it I can change to decreasing True or False.

##Define some useful colors

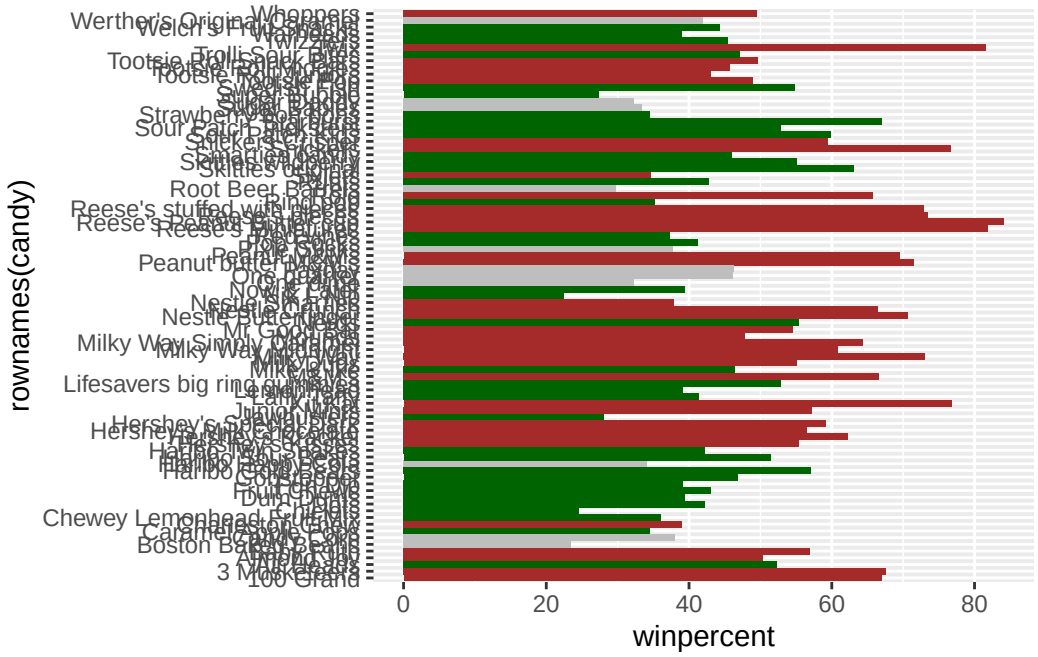
```
mycols <- rep("gray", nrow(candy))
mycols[as.logical(candy$fruity)] <- "darkgreen"
mycols[as.logical(candy$chocolate)] <- "brown"

mycols
```

```
 [1] "brown"     "brown"     "gray"      "gray"      "darkgreen" "brown"
 [7] "brown"     "gray"      "gray"      "darkgreen" "brown"     "darkgreen"
```

```
 [13] "darkgreen" "darkgreen" "darkgreen" "darkgreen" "darkgreen" "darkgreen"
 [19] "darkgreen" "gray"      "darkgreen" "darkgreen" "brown"     "brown"
 [25] "brown"     "brown"     "darkgreen" "brown"     "brown"     "darkgreen"
 [31] "darkgreen" "darkgreen" "brown"     "brown"     "darkgreen" "brown"
 [37] "brown"     "brown"     "brown"     "brown"     "brown"     "darkgreen"
 [43] "brown"     "brown"     "darkgreen" "darkgreen" "gray"      "brown"
 [49] "gray"      "darkgreen" "darkgreen" "brown"     "brown"     "brown"
 [55] "brown"     "darkgreen" "brown"     "gray"      "darkgreen" "brown"
 [61] "darkgreen" "darkgreen" "brown"     "darkgreen" "brown"     "brown"
 [67] "darkgreen" "darkgreen" "darkgreen" "darkgreen" "gray"      "gray"
 [73] "darkgreen" "darkgreen" "brown"     "brown"     "brown"     "brown"
 [79] "darkgreen" "brown"     "darkgreen" "darkgreen" "darkgreen" "gray"
 [85] "brown"
```

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

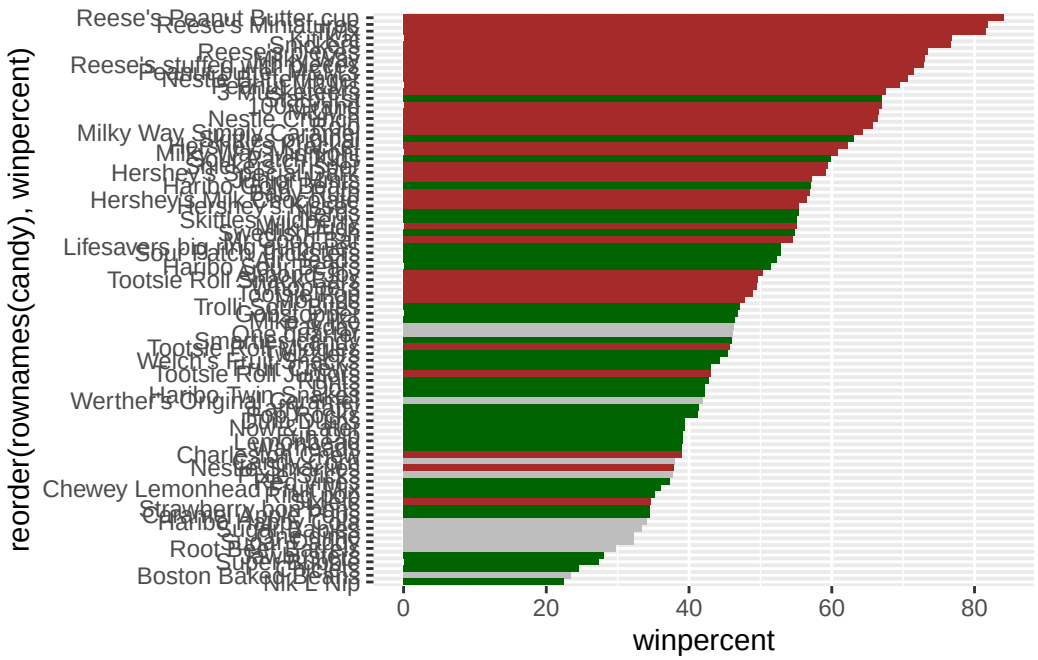
```
library(ggplot2)

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




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(fill=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 chocholate candy

Q18. What is the best ranked fruity candy?

The worst ranked fruity candy is Nik L nip

##4. Taking a look at pricepercent

```
library(ggrepel)

# 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=2.0, max.overlaps = 5)
```

```
Warning: ggrepel: 50 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 miniature as the best option regarding price percent and winpercent.

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

```
expensive <- order(candy$pricepercent, decreasing=TRUE)

head(candy[expensive,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
```

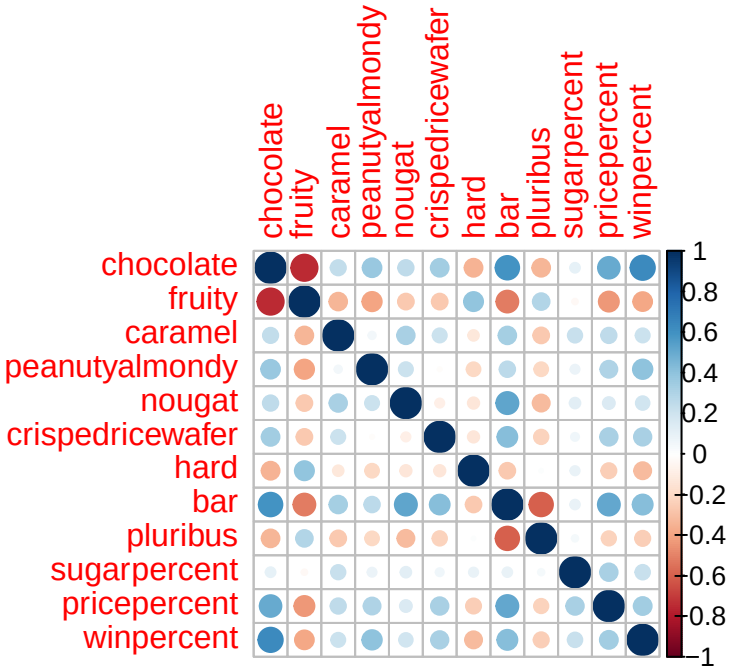
The most expensive is the least popular, Nik L Nip

##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)?

Two of the variables with the strongest anti-correlation in this plot are, being chocolate and being fruity, as well as being a bar and coming in multiple amounts (pluribus).

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

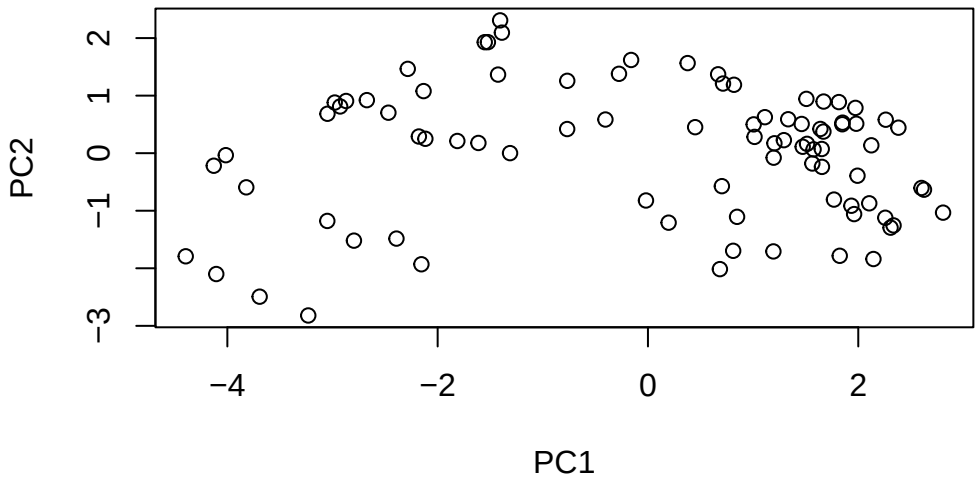
The most positively correlated variables are being a chocolate and win percentage is usually correlated, as well as a being a choocolate and being a bar.

##6. Principal Component Analysis

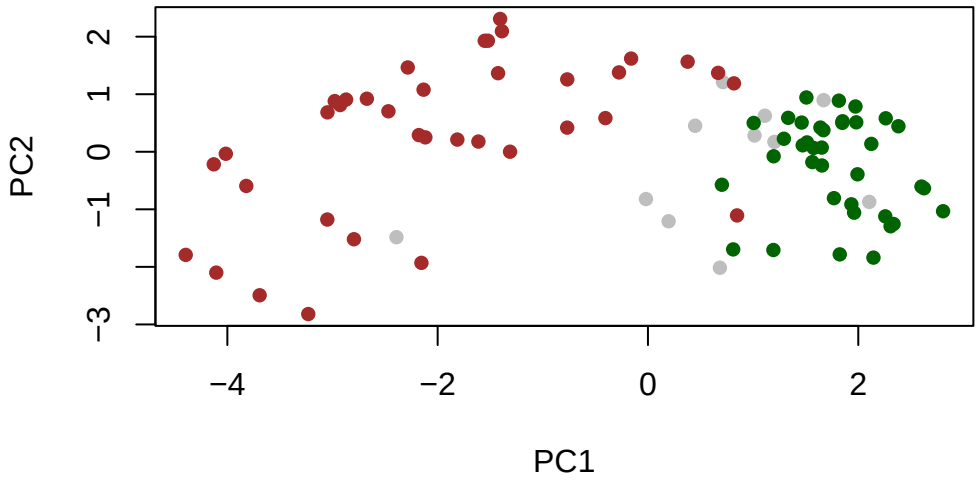
```
pca<- prcomp(candy, scale=TRUE)

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])
```

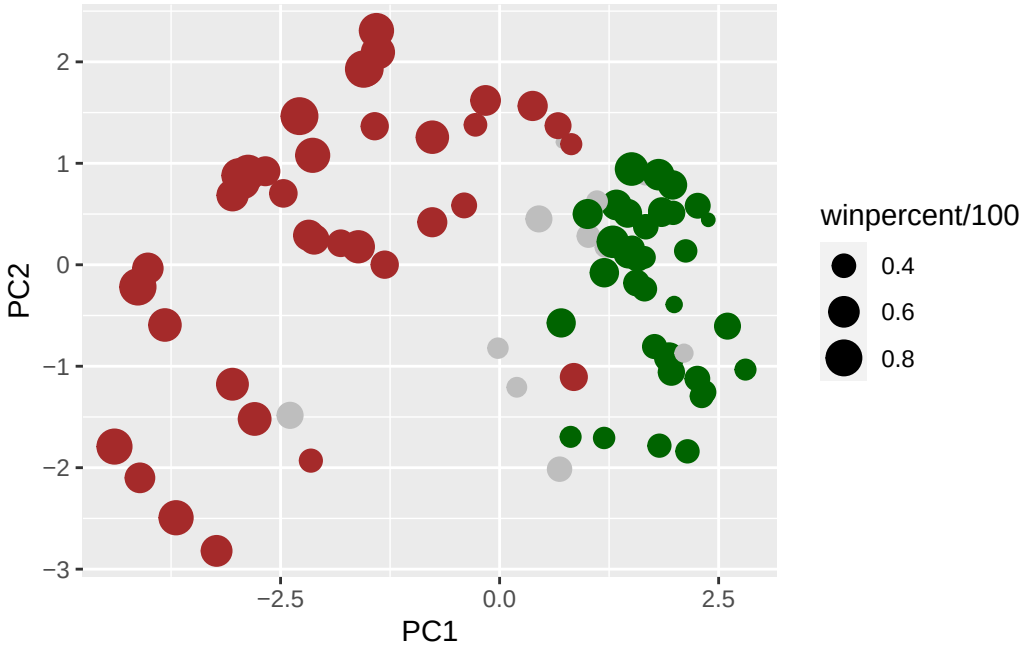


```
plot(pca$x[,1:2], col=mycols, pch=16)
```



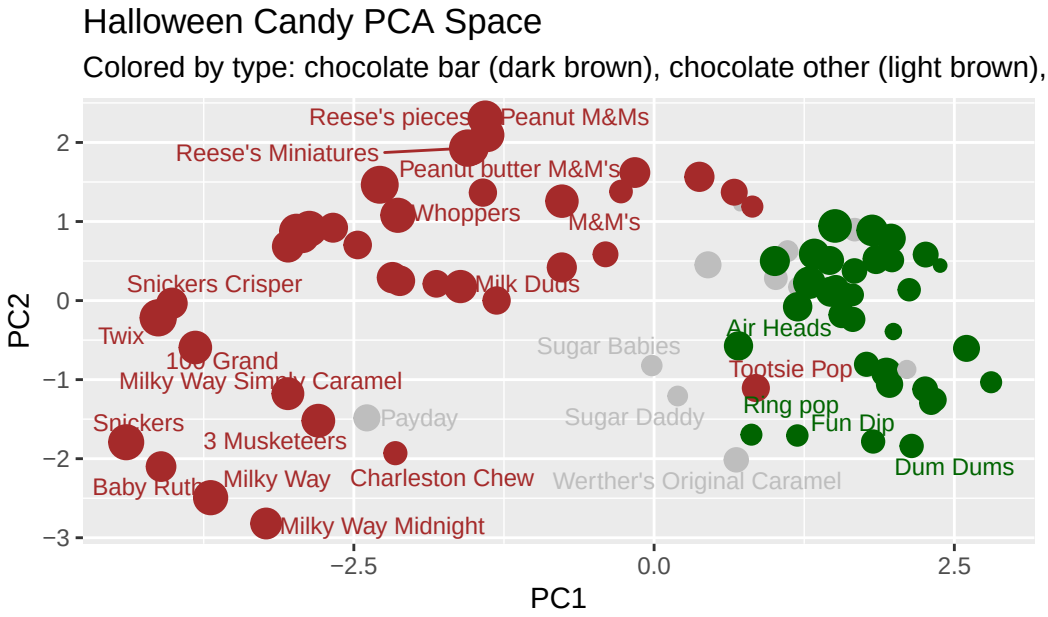

```
my_data <- cbind(candy, pca$x[,1:3])
```

```
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=mycols)

p
```

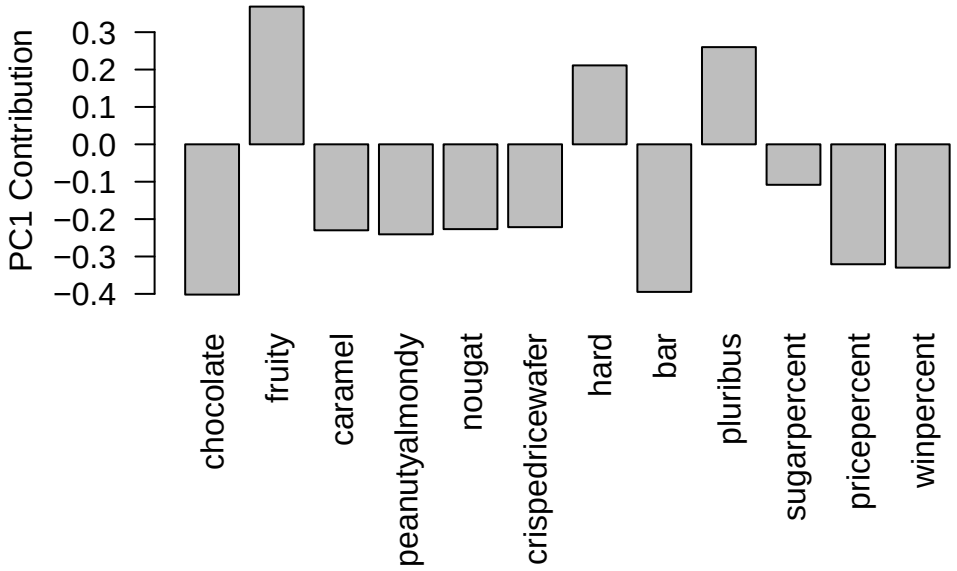



```
library(ggrepel)

p + geom_text_repel(size=3.3, col=mycols, max.overlaps = 7)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
       caption="Data from 538")
```

```
Warning: ggrepel: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




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



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

The variables in the correlation that were positively correlated with each other, therefore, it is makes sense, the negative direction are also correlated to each other but mostly in chocolate, for positive PC1, it is fruity candies that are usually hard and come in a bag or box of multiple candies.