class 10: halloween

Mia Fava

```
library(knitr)
```

##Importing Candy Data

```
candy_file <- "class10.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?

```
num_candies <- nrow(candy)
print(num_candies)
```

```
[1] 85
```

85 Different types

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

```
table(candy$fruity)
```

```
 0  1
47 38
```

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

There are 38 fruity candies

##What's your favorite candy?

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

```
[1] 81.64291
```

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

```
favorite_candy <- "Swedish Fish"
favorite_winpercent <- candy[favorite_candy, "winpercent"]
print(favorite_winpercent)
```

```
[1] 54.86111
```

My favorite candy is Swedish Fish with a winpercent of 54.9%

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

```
kitkat_winpercent <- candy["Kit Kat", "winpercent"]
print(kitkat_winpercent)
```

```
[1] 76.7686
```

Winpercent of Kit Kat is 76.8%

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

```
tootsie_winpercent <- candy["Tootsie Roll Snack Bars", "winpercent"]
print(tootsie_winpercent)
```

```
[1] 49.6535
```

Winpercent of Toosie Roll Snack Bars is 49.7%

> Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

```
library(skimr)
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	

The column that seems to be on a different scale is likely the winpercent due to how it is measured as a percentage. The other columns are measured as binary values, while winpercent has a range of values that are seen to be continuous.

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

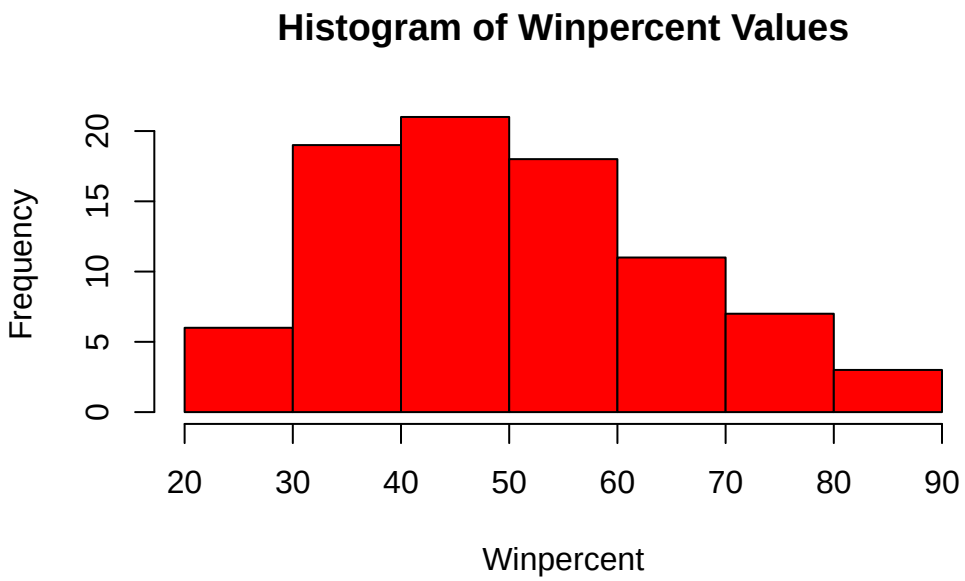
```
unique(candy$chocolate)
```

```
[1] 1 0
```

For this column, the 0 infers that the candy does not contain chocolate and a 1 means that the candy contains chocolate.

Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent,
     main = "Histogram of Winpercent Values",
     xlab = "Winpercent",
     col = "red",
     border = "black")
```



Q9. Is the distribution of winpercent values symmetrical?

The histrgram shows that the distribution of winpercent are not perfectly symmetrical, there is a slight skew to the left as seen in the graph above.

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

```
median_winpercent <- median(candy$winpercent)
print(median_winpercent)
```

```
[1] 47.82975
```

The center is the at 48%.

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

```
chocolate_winpercent <- mean(candy$winpercent[as.logical(candy$chocolate)])
print(chocolate_winpercent)
```

```
[1] 60.92153
```

```
fruity_winpercent <- mean(candy$winpercent[as.logical(candy$fruity)])
print(fruity_winpercent)
```

```
[1] 44.11974
```

Chocolate is ranked higher as $60.92\% > 44.12\%$.

Q12. Is this difference statistically significant?

```
t_test_result <- t.test(candy$winpercent[as.logical(candy$chocolate)],
print(t_test_result)
```

```
    Welch Two Sample t-test

data:  candy$winpercent[as.logical(candy$chocolate)] and candy$winpercent[as.logical(candy$fi
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
```

The p value is less than 0.05, so the difference is statisically significant

##Overall Candy Ranking

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

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

The 5 least liked candies Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbuster.

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

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

```
                        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 favorite candies are Reese's Peanut Butter Cup, Reese's Miniatures, Twix, Kit Kat, Snickers.

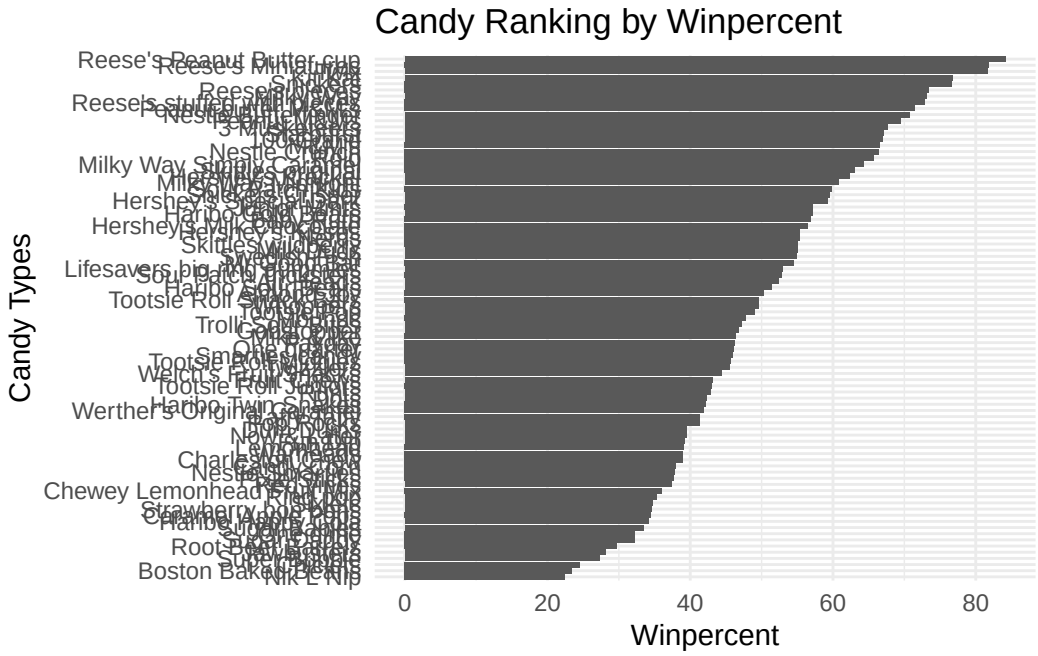
Q15. Make a first barplot of candy ranking based on winpercent values.

```
if (!require("ggplot2")) install.packages("ggplot2", dependencies = TRUE)
```

```
Loading required package: ggplot2
```

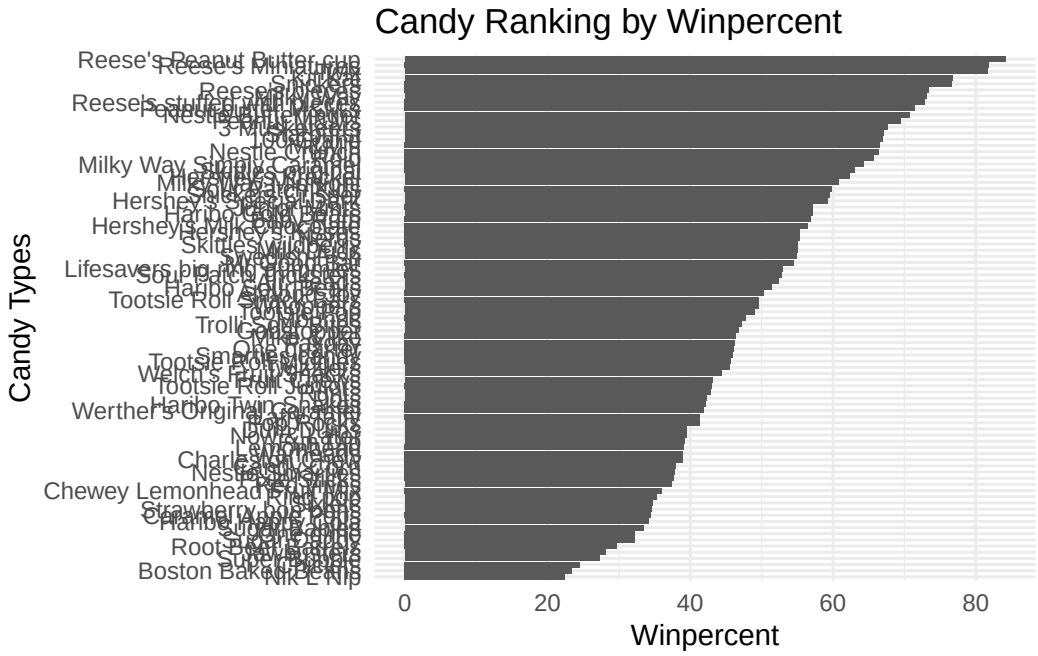
```
library(ggplot2)
```

```
ggplot(candy, aes(x = winpercent, y = reorder(rownames(candy), winpercent))) +
  geom_bar(stat = "identity") +
  labs(title = "Candy Ranking by Winpercent",
       x = "Winpercent",
       y = "Candy Types") +
  theme_minimal()
```



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

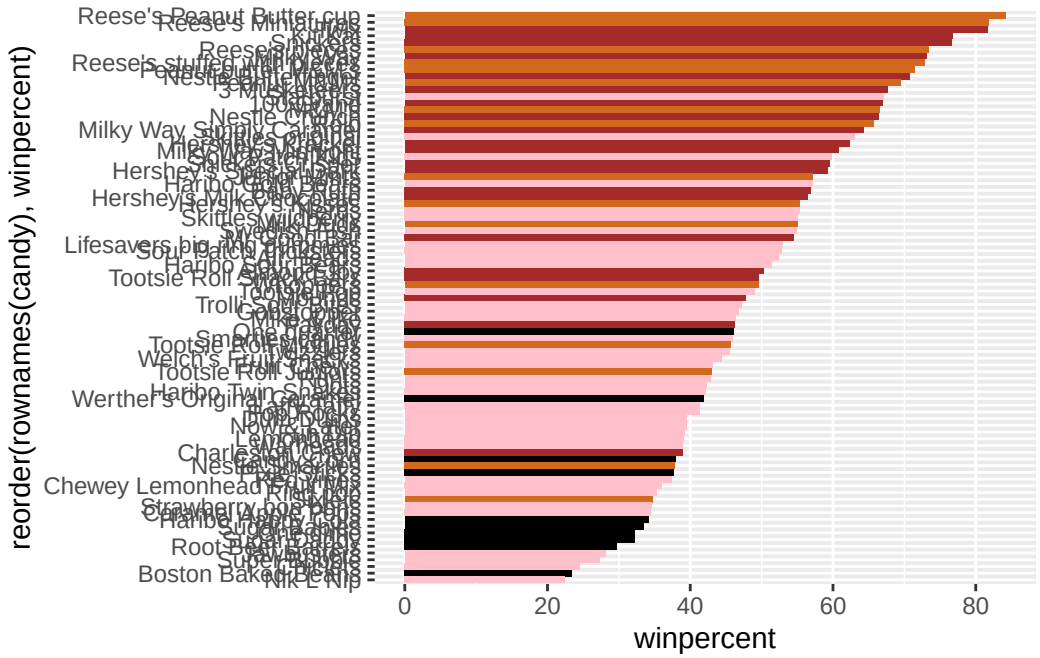
```
ggplot(candy, aes(x = winpercent, y = reorder(rownames(candy), winpercent))) +
  geom_bar(stat = "identity") +
  labs(title = "Candy Ranking by Winpercent",
       x = "Winpercent",
       y = "Candy Types") +
  theme_minimal()
```



##Add 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?

```
worst_chocolate <- candy[candy$chocolate == 1, ]
worst_chocolate <-worst_chocolate[order(worst_chocolate$winpercent), ]
worst_chocolate <- head(worst_chocolate, 1)
print(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
```

The worst chocolate are Sixlets.

Q18. What is the best ranked fruity candy?

```
best_fruity <- candy[candy$fruity == 1, ]
best_fruity <-best_fruity[order(-best_fruity$winpercent), ]
best_fruity <- head(best_fruity, 1)
print(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
```

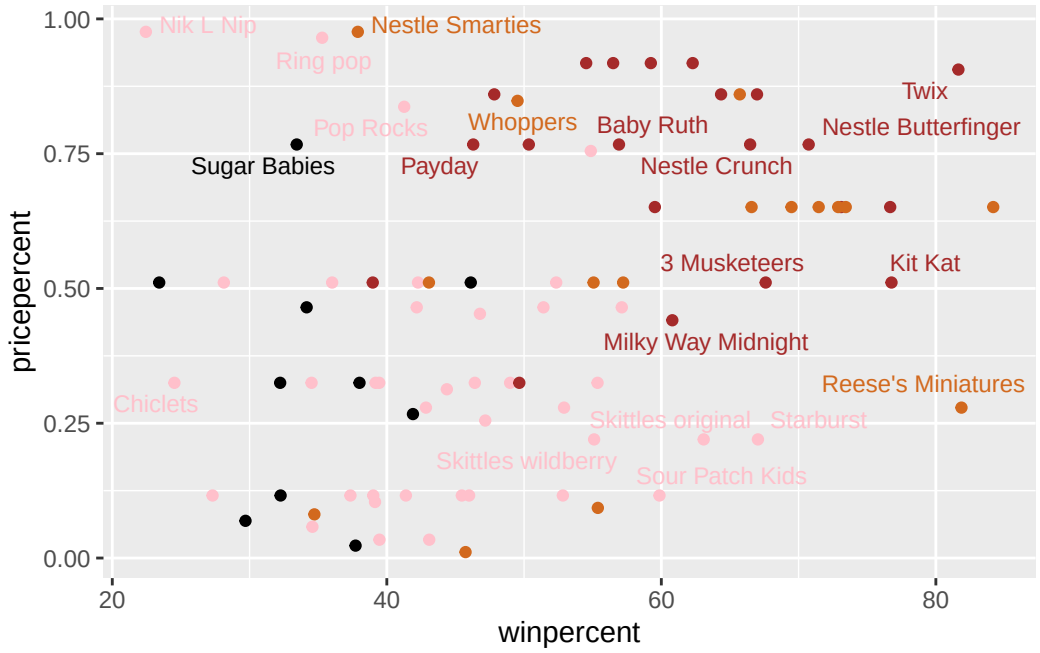
The best fruity candy are Starbursts.

##Taking a look at pricepercent:

```
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?

```
candy$bang_for_buck <- candy$winpercent / candy$pricepercent
best_value_candy <- candy[which.max(candy$bang_for_buck), ]
print(best_value_candy)
```

```
                     chocolate fruity caramel peanutyalmondy nougat
Tootsie Roll Midgies         1      0       0              0      0
                     crispedricewafer hard bar pluribus sugarpercent
Tootsie Roll Midgies                0    0   0        1        0.174
                     pricepercent winpercent bang_for_buck
Tootsie Roll Midgies        0.011   45.73675      4157.886
```

The best candy that offers the most bang for your buck are the Tootsie Roll Midgies.

> 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)
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
```

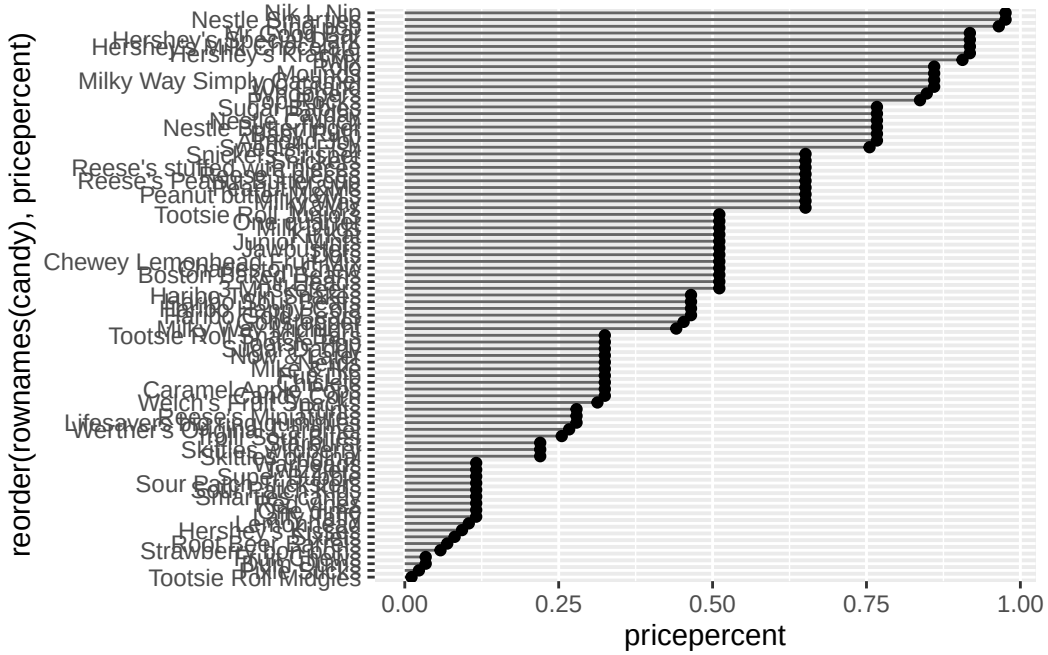
```
top_expensive <- candy[order(-candy$pricepercent), ][1:5, ]
least_popular_expensive <- top_expensive[which.min(top_expensive$winpercent), ]
print(least_popular_expensive)
```

```
          chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
Nik L Nip         0      1       0              0      0                0    0
          bar pluribus sugarpercent pricepercent winpercent bang_for_buck
Nik L Nip   0        1        0.197        0.976   22.44534      22.99728
```

The most expensive candies and the least populat among them are the Nik L Nip.

> 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()

```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```

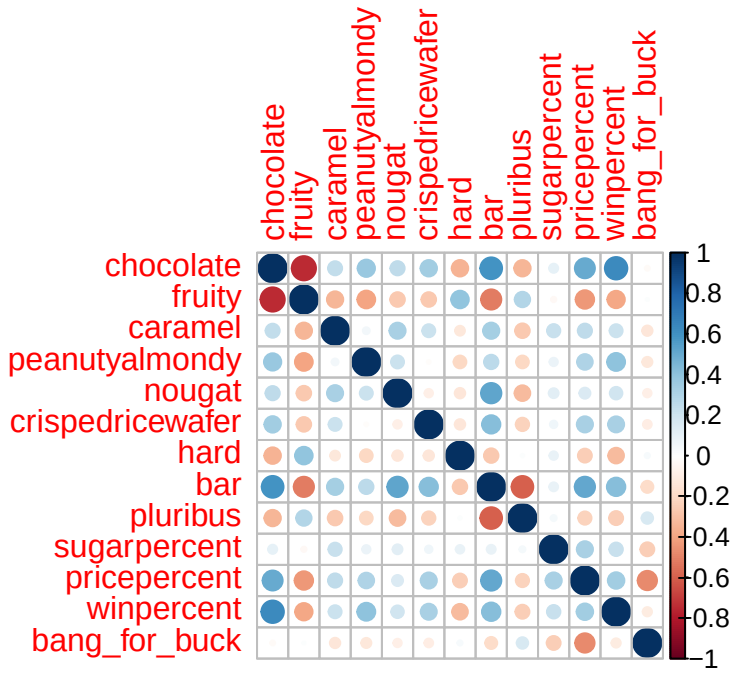


##Exploring the correlation structure

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



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

```
anti_correlated <- which(cij < -0.5, arr.ind = TRUE)
anti_correlated_vars <- data.frame(
    Var1 = rownames(cij)[anti_correlated[, 1]],
    Var2 = colnames(cij)[anti_correlated[, 2]],
    Correlation = cij[anti_correlated]
)
print(anti_correlated_vars)
```

```
       Var1      Var2 Correlation
1    fruity chocolate  -0.7417211
2 chocolate    fruity  -0.7417211
3       bar    fruity  -0.5150656
4    fruity       bar  -0.5150656
5  pluribus       bar  -0.5934089
6       bar  pluribus  -0.5934089
```

The two variables that are anti-correlated are fruity and chocolate or fruity and bar, or bar and pluribus.

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

```r
positive_correlated <- which(cij > 0.5, arr.ind = TRUE)
positive_correlated_vars <- data.frame(
    Var1 = rownames(cij)[positive_correlated[, 1]],
    Var2 = colnames(cij)[positive_correlated[, 2]],
    Correlation = cij[positive_correlated]
)
print(positive_correlated_vars)
```

```
               Var1             Var2 Correlation
1         chocolate        chocolate   1.0000000
2               bar        chocolate   0.5974211
3      pricepercent        chocolate   0.5046754
4        winpercent        chocolate   0.6365167
5            fruity           fruity   1.0000000
6           caramel          caramel   1.0000000
7    peanutyalmondy   peanutyalmondy   1.0000000
8            nougat           nougat   1.0000000
9               bar           nougat   0.5229764
10 crispedricewafer crispedricewafer   1.0000000
11             hard             hard   1.0000000
12        chocolate              bar   0.5974211
13           nougat              bar   0.5229764
14              bar              bar   1.0000000
15     pricepercent              bar   0.5184065
16         pluribus         pluribus   1.0000000
17     sugarpercent     sugarpercent   1.0000000
18        chocolate     pricepercent   0.5046754
19              bar     pricepercent   0.5184065
20     pricepercent     pricepercent   1.0000000
21        chocolate       winpercent   0.6365167
22       winpercent       winpercent   1.0000000
23    bang_for_buck    bang_for_buck   1.0000000
```

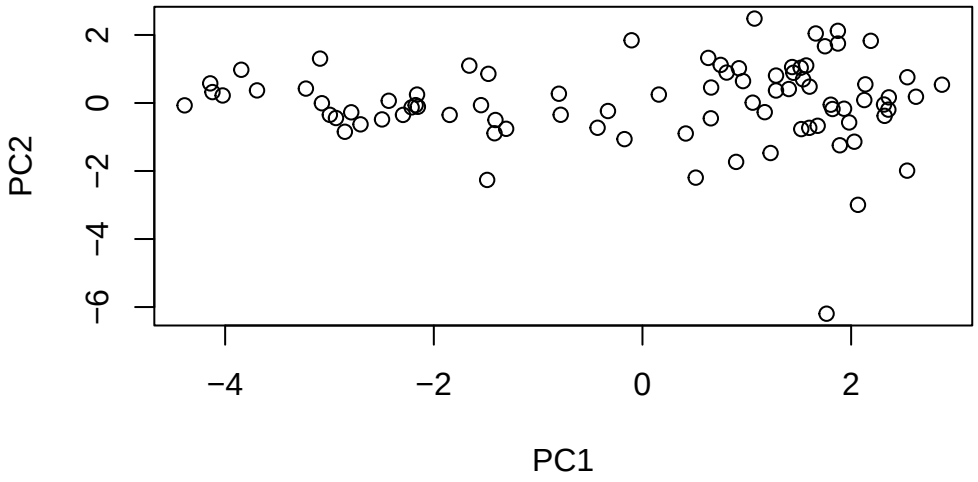
The variables that are most positively correlated are when Var 1 = Var 2, so the correlation is 1 – for example, chocolate and chocolate have 1 correlation.

##PCA

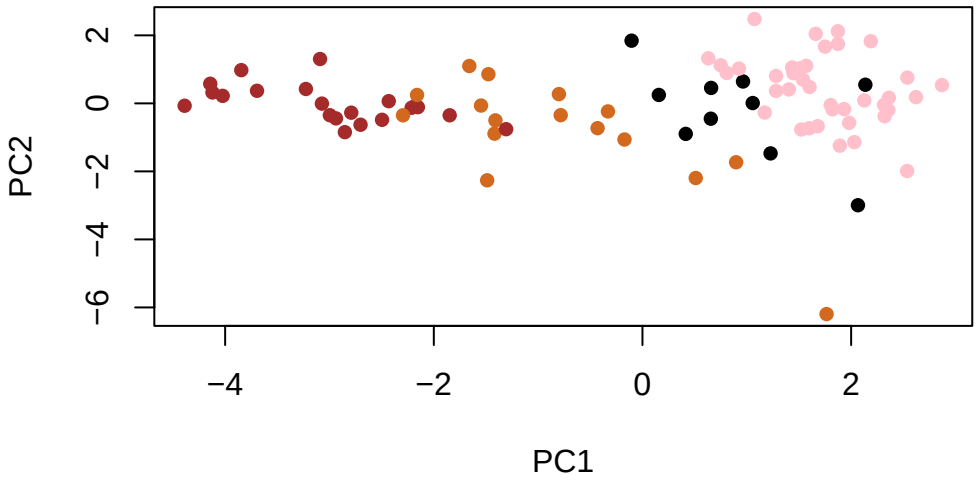
```r
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7
Standard deviation     2.0938 1.2127 1.13054 1.0787 0.98027 0.93656 0.81530
Proportion of Variance 0.3372 0.1131 0.09832 0.0895 0.07392 0.06747 0.05113
Cumulative Proportion  0.3372 0.4503 0.54866 0.6382 0.71208 0.77956 0.83069
                          PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     0.78462 0.68466 0.66328 0.57829 0.43128 0.39534
Proportion of Variance 0.04736 0.03606 0.03384 0.02572 0.01431 0.01202
Cumulative Proportion  0.87804 0.91410 0.94794 0.97367 0.98798 1.00000
```

```r
plot(pca$x[, 1:2],
     xlab = "PC1",
     ylab = "PC2")
```

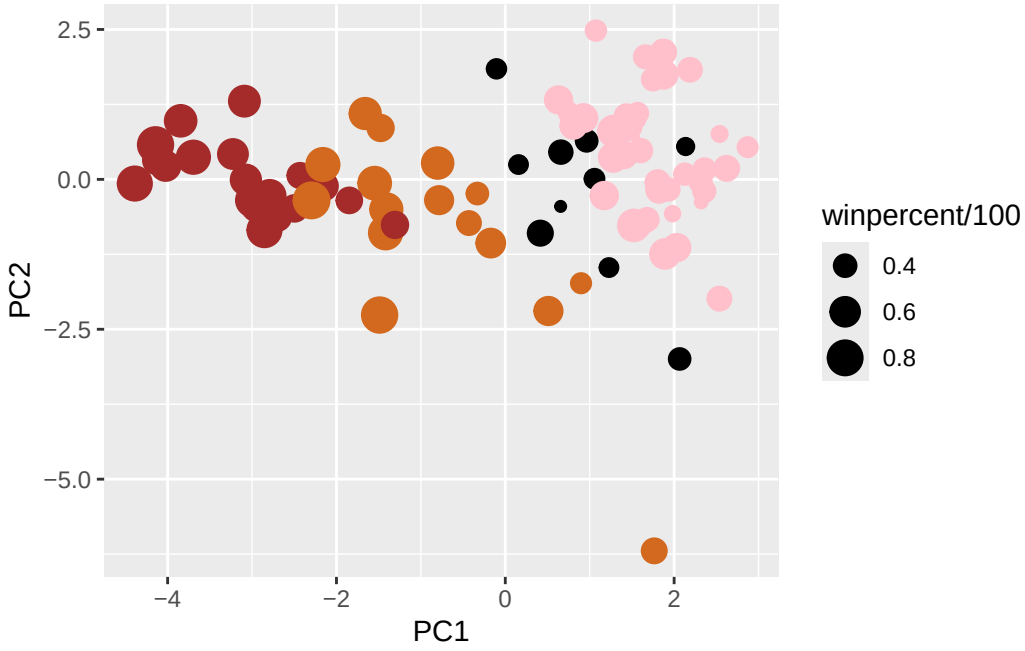



```r
plot(pca$x[,1:2], col=my_cols, pch=16)
```

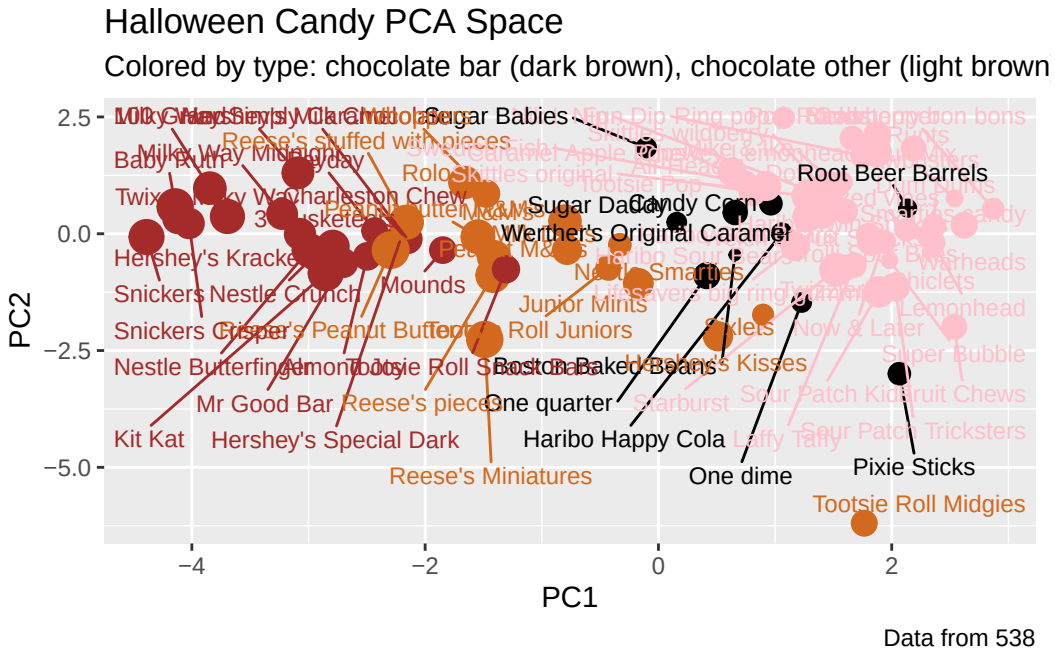


```
my_data <- cbind(candy, pca$x[,1:3])
```

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

p
```




```
library(ggrepel)

p + geom_text_repel(size=3, col=my_cols, max.overlaps = 50)  +
  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")
```




```
library(plotly)
```

```
Attaching package: 'plotly'
```

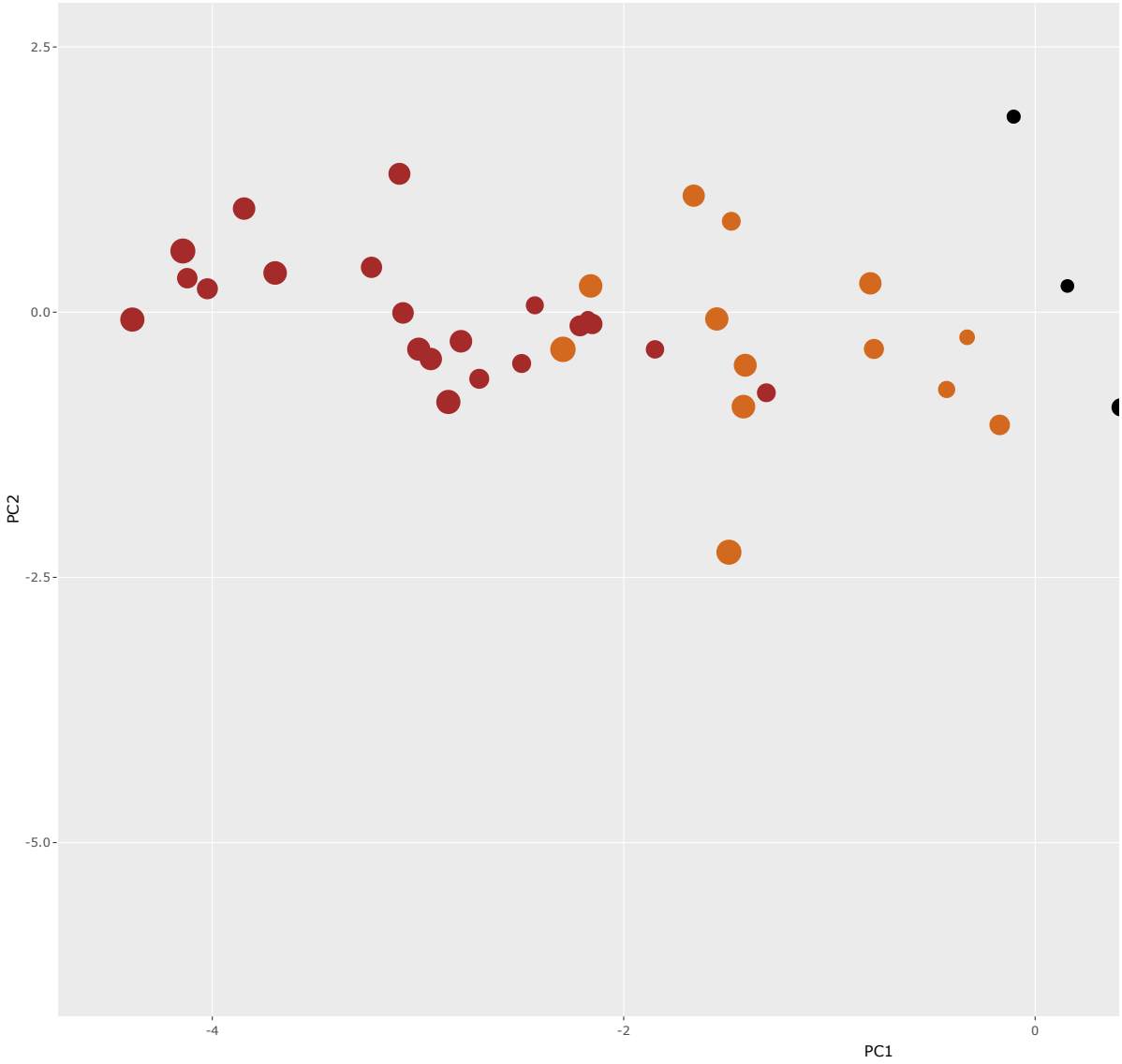
```
The following object is masked from 'package:ggplot2':

    last_plot
```

```
The following object is masked from 'package:stats':

    filter
```

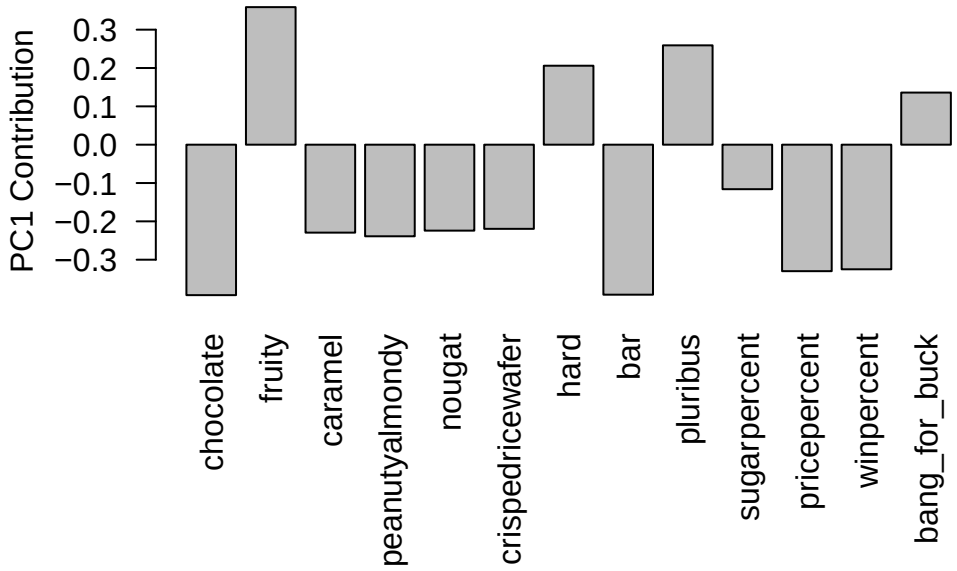
```
The following object is masked from 'package:graphics':

    layout
```

```
ggplotly(p)
```

```
PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installe
```




```
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?

```
pca$rotation[,1]
```

```
      chocolate           fruity          caramel   peanutyalmondy
     -0.3924439        0.3588085       -0.2293954       -0.2389173
         nougat crispedricewafer             hard              bar
     -0.2241826       -0.2195121        0.2059573       -0.3912663
        pluribus     sugarpercent     pricepercent       winpercent
      0.2590791       -0.1161206       -0.3299041       -0.3250778
bang_for_buck
      0.1359085
```

This does make sense with what the chart is showing due to how all the positive numbers related to the fruity, hard, pluribus, and bang for buck categories – this is evident in both the numerical and bar graph data. These 4 categories are seen to be positive for PC1.